TRANSFER LEARNING VIA UNSUPERVISED TASK DISCOVERY FOR VISUAL QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

We study how to leverage off-the-shelf visual and linguistic data to cope with out-of-vocabulary answers in visual question answering. Existing large-scale visual data with annotations such as image class labels, bounding boxes and region descriptions are good sources for learning rich and diverse visual concepts. However, it is not straightforward how the visual concepts should be captured and transferred to visual question answering models due to missing link between question dependent answering models and visual data without question or task specification. We tackle this problem in two steps: 1) learning a task conditional visual classifier based on unsupervised task discovery and 2) transferring and adapting the task conditional visual classifier to visual question answering models. Specifically, we employ linguistic knowledge sources such as structured lexical database (*e.g.* Wordnet) and visual descriptions for unsupervised task discovery, and adapt a learned task conditional visual classifier to answering unit in a visual question answering model. We empirically show that the proposed algorithm generalizes to unseen answers successfully using the knowledge transferred from the visual data.

1 INTRODUCTION

People see and understand a visual scene from diverse perspectives. For example, from a single image of a chair, people effortlessly recognize diverse visual concepts such as color, material, style, usage, and so on. These diverse perspectives may be associated with different questions in natural language for which people find appropriate answers. Recently visual question answering (VQA) (Antol et al., 2015) is proposed as an effort to learn deep neural network models with capability to perform diverse visual recognition tasks defined adaptively by questions.

Approaches to VQA rely on a large-scale dataset of image, question and answer triples, and train a classifier taking an image and a question as inputs and producing an answer. Despite recent remarkable progress (Yang et al., 2016; Fukui et al., 2016; Anderson et al., 2018), this direction has a critical limitation that image, question and answer triples in datasets are the only source for learning visual concepts. Such drawback may result in lack of scalability because the triplets may be collected artificially by human annotators with limited quality control and have weak diversity in visual concepts. In fact, VQA datasets (Goyal et al., 2017; Agrawal et al., 2018) suffer from inherent bias, which hinders learning true visual concepts from the datasets. On the contrary, people answer a question based on visual concepts learned from diverse sources such as personal experience, books, pictures, and videos, which are not necessarily associated with target questions. Even for machines, there exist more natural and scalable sources for learning visual concepts: image class labels, bounding boxes and region descriptions. Such information is already available in large-scale (Deng et al., 2009; Krasin et al., 2017; Krishna et al., 2017) and can scale further with reasonable cost (Papadopoulos et al., 2016; 2017). This observation brings a natural question; can we learn visual concepts without question annotations and transfer them for VQA?

To address this question, we introduce VQA with out-of-vocabulary answers, which is illustrated in Figure 1. External visual data provide a set of labels $\mathcal{A}$ and only a subset of these labels $\mathcal{B} \subset \mathcal{A}$ appears in VQA training set as answers. The goal of this setting is to handle out-of-vocabulary answers $\mathbf{a} \in \mathcal{A} - \mathcal{B}$ successfully by exploiting visual concepts learned from external visual data. To address this problem, this paper studies how to learn visual concepts without questions and how to transfer the learned concepts to VQA models. To learn transferable visual concepts, we train a

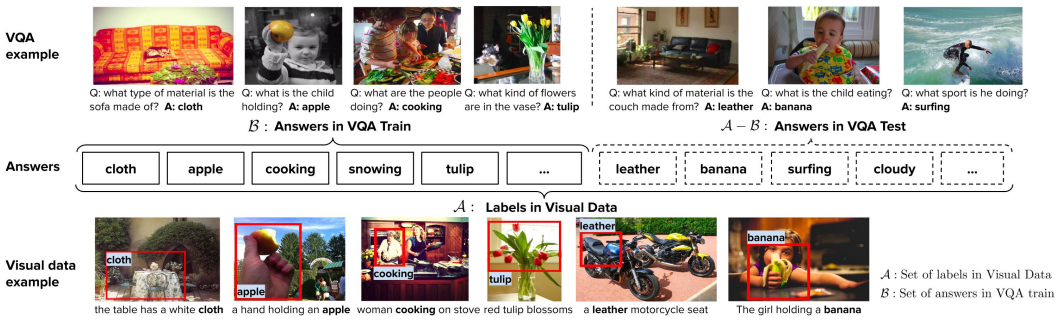


Figure 1: **VQA with out-of-vocabulary answers.** Given a set of labels in visual data $\mathcal{A}$ and a set of answers in VQA training set $\mathcal{B}$, we evaluate the model on VQA test set with answers $\mathbf{a} \in \mathcal{A} - \mathcal{B}$. External visual data provide a set of bounding box labels and visual descriptions.

task conditional visual classifier, whose task is defined by a task specification vector. The classifier is used as an answering unit where a task specification vector is inferred from a question. To train the task conditional visual classifier without task annotations, we propose an unsupervised task discovery technique based on linguistic knowledge sources such as structured lexical databases, *e.g.*, Wordnet (Fellbaum, 1998), and region descriptions. We present that the proposed transfer learning helps generalization in VQA with out-of-vocabulary answers.

The main contribution of our paper is three-fold:

- We present a novel transfer learning algorithm for visual question answering based on a task conditional visual classifier.
- We propose an unsupervised task discovery technique for learning task conditional visual classifiers without explicit task annotations.
- We demonstrate that the proposed method enables to answer with out-of-vocabulary answers based on knowledge transfer from visual data without question annotations.

The rest of the paper is organized as follows. Section 2 discusses prior works related to our approach. We describe the overall transfer learning framework in Section 3. Learning visual concepts by unsupervised task discovery is described in Section 4. Section 5 analyzes experimental results and Section 6 makes our conclusion.

2 Related Works

Standard VQA evaluation assumes identically distributed train and test set (Malinowski & Fritz, 2014; Antol et al., 2015; Zhu et al., 2016). As this evaluation setting turns out to be vulnerable to models exploiting biases in training set (Goyal et al., 2017), several alternatives have been proposed. One approach is to reduce observed biases either by balancing answers for individual questions (Goyal et al., 2017) or by providing different biases to train and test sets intentionally (Agrawal et al., 2018). Another approach is to construct compositional generalization split (Johnson et al., 2017; Agrawal et al., 2017) whose question and answer pairs in test set are formed by novel compositions of visual concepts and question types appearing in the training set. This split is constructed by repurposing an existing VQA dataset (Agrawal et al., 2017) or by constructing a synthetic dataset (Johnson et al., 2017). The problem setting studied in this paper is similar to Teney & Hengel (2016) in the sense that out-of-vocabulary answers are used for testing, but unlike the prior work, we formulate the problem as a transfer learning where out-of-vocabulary answers are learned from external visual data.

External data are often employed in VQA for better generalization. Convolutional neural networks (Krizhevsky et al., 2012; He et al., 2016) pretrained on ImageNet (Deng et al., 2009) is a widely accepted standard for diverse VQA models (Yang et al., 2016; Fukui et al., 2016). As an alternative, object detector (Ren et al., 2015) trained on Visual Genome (Krishna et al., 2017) is employed to extract pretrained visual features (Anderson et al., 2018). Pretrained language models such as word embeddings (Pennington et al., 2014) or sentence embeddings (Kiros et al., 2015) are frequently used

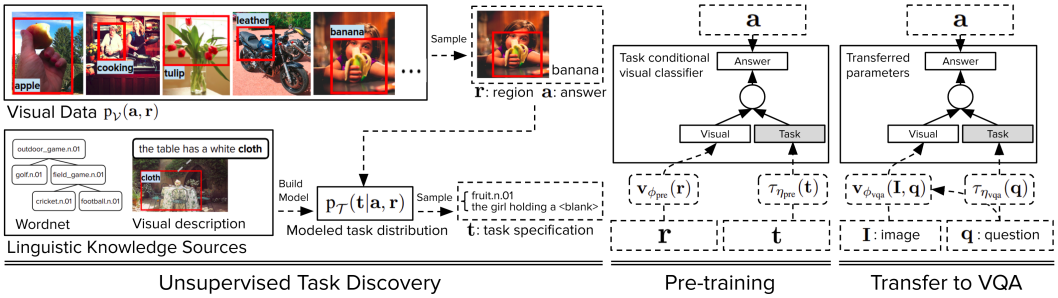


Figure 2: **Overview of the proposed algorithm.** (Left) Unsupervised task discovery learns a task conditional visual classifier by leveraging off-the-shelf visual data without task specification $\mathbf{t}$. It also defines a task distribution $\mathrm{p}_{\mathcal{T}}(\mathbf{t}|\mathbf{a}, \mathbf{r})$ using linguistic knowledge sources, where stochastic sampling associates a task specification $\mathbf{t}$ with a visual annotation $(\mathbf{a}, \mathbf{r})$. (Center) A visual annotation with a task specification, denoted by $(\mathbf{a}, \mathbf{r}, \mathbf{t})$, is employed to pretrain a task conditional visual classifier. (Right) Pretrained task conditional visual classifier is transferred to VQA with the learned parameters and adapts input representations $\mathbf{v}_{\phi_{\text{vqa}}}(\mathbf{I}, \mathbf{q})$ and $\tau_{\eta_{\text{vqa}}}(\mathbf{q})$ without fine-tuning.

to initialize parameters of question encoders (Noh et al., 2016; Fukui et al., 2016; Teney et al., 2018). Exploiting information retrieval from knowledge base (Auer et al., 2007; Bollacker et al., 2008) or external vision algorithms (Wang et al., 2017b) to provide additional inputs to VQA models was investigated in (Wu et al., 2016; Wang et al., 2017a;b). Sharing aligned image-word representations between VQA models and image classifiers was proposed in (Gupta et al., 2017) to exploit external visual data. However, transfer learning from external data to cope with out-of-vocabulary answers in VQA has hardly been investigated so far.

Transfer learning from external data to cope with out-of-vocabulary words is actively studied in novel object captioning (Anne Hendricks et al., 2016; Venugopalan et al., 2017; Yao et al., 2017; Lu et al., 2018). Anne Hendricks et al. (2016) and Venugopalan et al. (2017) decompose image captioning into visual classification and language modeling and exploit unpaired visual and linguistic data as additional resources to train visual classifier and language model respectively. Recent approaches incorporate pointer networks (Vinyals et al., 2015) and learn to point an index of word candidates (Yao et al., 2017) or an associated region (Lu et al., 2018), where the word candidates are detected by a multi-label classifier (Yao et al., 2017) or an object detector (Lu et al., 2018) trained with external visual data. These algorithms are not directly applicable to our problem setting because they focus on predicting object words without task specification while the task conditional visual recognition is required for VQA.

3 Algorithm Overview

The main objective of this paper is to learn visual concepts using off-the-shelf visual data and transfer the concepts to VQA models. To adapt the learned visual concepts to diverse questions about a visual scene, we should learn not only a name of a visual concept but also a type of the concept. For example, a question about a dog image can be about name of the animal as well as its visual attributes such as breed and color. We introduce a task conditional visual classifier for the purpose, in which a type or granularity of the concept is specified by a task specification vector and the final answer is an output of the specified recognition task. We pretrain this task conditional visual classifier using visual data without question or task specifications, and adapt it to VQA models by transferring the learned parameters. Figure 2 illustrates overview of the proposed algorithm.

3.1 Task Conditional Visual Classifier

Task conditional visual classifier is a function taking a visual feature $\mathbf{v} \in \mathbb{R}^d$ and a task specification vector $\tau \in \mathbb{R}^k$ and producing a probability distribution of answers $\mathbf{a} \in [0, 1]^l$. It is formulated as a neural network with parameter θ, and models a conditional distribution $\mathrm{p}_\theta(\mathbf{a}|\mathbf{v}, \tau)$. The inputs $\mathbf{v}$ and τ are typically constructed by external feature encoders $\mathbf{v}_\phi(\cdot)$ and $\tau_\eta(\cdot)$, respectively.

In the proposed transfer learning scenario, a task conditional visual classifier is pretrained with off-the-shelf visual data and transferred to VQA. In the pretraining stage, the model parameter θ and the parameters for external feature encoders ϕ and η are jointly learned by back-propagation as described in Equation 2. This stage allows a task conditional visual classifier to learn diverse visual recognition tasks $\mathrm{p}_\tau(\mathbf{a}|\mathbf{v})$ by varying task specification vector τ. Transfer learning to VQA is achieved by reusing parameter θ and adapting another external feature encoders $\mathbf{v}_{\phi_{\text{vqa}}}(\cdot)$ and $\tau_{\eta_{\text{vqa}}}(\cdot)$ to task conditional visual classifier. We first describe transfer learning to VQA in Section 3.2 while pretraining a task conditional visual classifier with off-the-self visual data by unsupervised task discovery is described in Section 4.

3.2 Transfer Learning for Visual Question Answering

As illustrated in Figure 2, the proposed VQA model contains a task conditional visual classifier $\mathrm{p}_\theta(\mathbf{a}|\mathbf{v},\tau)$. The pretrained visual concepts are transferred to VQA by defining the function with the learned parameters θ. Then, learning a VQA model is now formulated as learning input representations $\mathbf{v}$ and τ for $\mathrm{p}_\theta(\mathbf{a}|\mathbf{v},\tau)$, which is given by

$$\phi^*_{\text{vqa}}, \eta^*_{\text{vqa}} = \underset{\phi_{\text{vqa}},\eta_{\text{vqa}}}{\arg\max}\, \mathbb{E}_{\mathrm{p}_{\text{vqa}}(\mathbf{a},\mathbf{I},\mathbf{q})} \log \mathrm{p}_\theta(\mathbf{a}|\mathbf{v}_{\phi_{\text{vqa}}}(\mathbf{I},\mathbf{q}), \tau_{\eta_{\text{vqa}}}(\mathbf{q})), \tag{1}$$

where $\mathbf{v}_{\phi_{\text{vqa}}}(\mathbf{I},\mathbf{q})$ is an encoded visual feature with an image $\mathbf{I}$ and a question $\mathbf{q}$ using an attention mechanism with parameter ϕ_{vqa}, and $\tau_{\eta_{\text{vqa}}}(\mathbf{q})$ is a task specification vector encoded with a question $\mathbf{q}$ using an encoder parameterized by η_{vqa}. The $\mathrm{p}_{\text{vqa}}(\mathbf{a},\mathbf{I},\mathbf{q})$ is a training data distribution of VQA dataset. We learn ϕ_{vqa} and η_{vqa} by stochastic gradient descent with maximum likelihood objective while the parameter for pretrained task conditional visual classifier θ remains fixed.

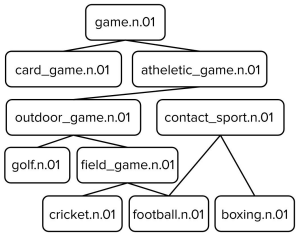
Matching visual features To reuse pretrained visual classifier in VQA without fine-tuning, visual features $\mathbf{v}$ should not be changed. This is fulfilled in recent approaches for VQA that do not fine-tune pretrained visual feature extractors. In this setting, all we need to do is using an identical visual feature extractor for both pretraining and VQA. Specifically, we use attention mechanism (Kim et al., 2016) on a pretrained bottom-up attention features (Anderson et al., 2018), where attention based on bounding boxes is used for pretraining and question-based attention is employed for VQA.

Weakly supervised task regression Utilizing a pretrained task conditional visual classifier to perform visual recognition specified by a question $\mathbf{q}$ requires to infer an task specification vector $\tau^*_{\mathbf{q}}$. This requirement introduces a learning problem—task regression—that optimizes an encoder $\tau_{\eta_{\text{vqa}}}(\mathbf{q})$ to correctly predict $\tau^*_{\mathbf{q}}$. Instead of directly minimizing $\mathcal{E}(\tau^*_{\mathbf{q}}, \tau_{\eta_{\text{vqa}}}(\mathbf{q}))$ with additional supervision, we exploit VQA data as a source of weak supervision. We propose to optimize a mapping from a visual feature to an answer using an indirect loss denoted by $\mathcal{E}(\mathrm{p}_{\tau^*_{\mathbf{q}}}(\mathbf{a}|\mathbf{v}), \mathrm{p}_\theta(\mathbf{a}|\mathbf{v}, \tau_{\eta_{\text{vqa}}}(\mathbf{q})))$; the resulting training objective is identical to Equation 1.

Out-of-vocabulary answering We learn VQA by adapting input representation while fixing the pretrained task conditional visual classifier $\mathrm{p}_\theta(\mathbf{a}|\mathbf{v},\tau)$. This strategy allows a model to focus on learning to infer a visual recognition task $\tau_{\eta_{\text{vqa}}}(\mathbf{q})$ from questions, which does not require data for all possible answers. Once the task specification vector τ is inferred, the learned task conditional visual classifier $\mathrm{p}_\theta(\mathbf{a}|\mathbf{v},\tau)$ can answer pretrained visual concepts including out-of-vocabulary answers.

4 Unsupervised Task Discovery

Learning a task conditional visual classifier with the off-the-shelf visual data (Krishna et al., 2017) is not straightforward due to missing annotation for task specifications. To address this issue, we propose unsupervised task discovery, which exploits a modeled task distribution to sample a task specification instead of collecting additional data. The motivation of this approach is that output domain of a visual recognition task reflects our understanding about hierarchy of concepts or knowledge of the world. For example, classification problems about *electronic devices* and *holdable objects* both restricts possible outputs to the corresponding word groups, which are defined by our prior knowledge. This knowledges is often accessible with linguistic knowledge sources, which is used for modeling a task distribution. We describe an algorithm to learn task conditional visual classifier with modeled task distribution in Section 4.1 and discuss how linguistic knowledge sources are used to model a task distribution in Section 4.2.



name $\mathbf{t}_s$	answer set $\mathcal{W}_{\mathbf{t}_s}$
appliance.n.02	freezer, hair dryer, refrigerator, fridge, oven, dishwasher, ..
opening.n.10	manhole, rear window, exit, nozzle, car window, coin slot, ..
food.n.02	potatoes, french fries, chicken, melted cheese, tomatoes, ..
move.v.03	twisted, ruffled, curving, curved, wheeled, rolled, coiled, ..
act.n.02	fishing, skateboarding, sailing, playing baseball, surfing, ..
area.n.01	middle, in corner, parking space, playground, landscape, ..
color.n.01	yellow, pink, red, beige, royal blue, amber, aqua, dark red, ..

Figure 3: **Illustration of Wordnet and extracted answer set.** (Left) A subgraph of the Wordnet (Fellbaum, 1998). Complex hierarchy of words reveals the diverse categorization of each words. (Right) A set of words sharing common parents in the tree is grouped as a single answer set. Diverse grouping of words reveals diverse level of understanding about the world.

4.1 Pretraining with Decomposed Data Distribution

Learning task conditional visual classifier is naturally formulated as maximizing expected log likelihood as

$$\theta^*, \phi^*_{\text{pre}}, \eta^*_{\text{pre}} = \underset{\theta, \phi_{\text{pre}}, \eta_{\text{pre}}}{\arg\max}\, \mathbb{E}_{\text{p}_{\mathcal{D}}(\mathbf{a},\mathbf{r},\mathbf{t})} \log \text{p}_\theta(\mathbf{a}|\mathbf{v}_{\phi_{\text{pre}}}(\mathbf{r}), \tau_{\eta_{\text{pre}}}(\mathbf{t})), \tag{2}$$

where $\mathbf{v}_{\phi_{\text{pre}}}(\mathbf{r})$ is a visual feature encoded by a region of interest $\mathbf{r}$ given by a whole image or a bounding box annotation, $\tau_{\eta_{\text{pre}}}(\mathbf{t})$ is a task specification vector regressed from a task specification $\mathbf{t}$, and $\{\theta, \phi_{\text{pre}}, \eta_{\text{pre}}\}$ are model parameters. This objective requires a joint distribution $\text{p}_{\mathcal{D}}(\mathbf{a}, \mathbf{r}, \mathbf{t})$, which is modeled by dense annotations of $(\mathbf{a}, \mathbf{r}, \mathbf{t})$ triples. However, this approach limits the utility of existing large scale visual annotations, which often consist of $(\mathbf{a}, \mathbf{r})$ pairs only.

We decompose the joint distribution into two components—one is for visual annotation $\text{p}_{\mathcal{V}}(\mathbf{a}, \mathbf{r})$ and the other is for task conditioned on visual annotation $\text{p}_{\mathcal{T}}(\mathbf{t}|\mathbf{a}, \mathbf{r})$—which is formally given by

$$\text{p}_{\mathcal{D}}(\mathbf{a}, \mathbf{r}, \mathbf{t}) = \text{p}_{\mathcal{T}}(\mathbf{t}|\mathbf{a}, \mathbf{r})\text{p}_{\mathcal{V}}(\mathbf{a}, \mathbf{r}). \tag{3}$$

Note that this formulation facilitates utilizing existing visual annotations $\text{p}_{\mathcal{V}}(\mathbf{a}, \mathbf{r})$ by appropriately modeling $\text{p}_{\mathcal{T}}(\mathbf{t}|\mathbf{a}, \mathbf{r})$. With the decomposed data distribution, the joint distribution for $(\mathbf{a}, \mathbf{r}, \mathbf{t})$ is approximated by sampling $(\mathbf{a}, \mathbf{r})$ from $\text{p}_{\mathcal{V}}(\mathbf{a}, \mathbf{r})$ followed by sampling $\mathbf{t}$ from $\text{p}_{\mathcal{T}}(\mathbf{t}|\mathbf{a}, \mathbf{r})$. Modeling $\text{p}_{\mathcal{T}}(\mathbf{t}|\mathbf{a}, \mathbf{r})$ with linguistic knowledge sources is described next.

4.2 Leveraging Linguistic Knowledge Sources

A visual recognition task defines a mapping from visual inputs to a set of answers, where a range of the mapping is a finite set of answers, which is a subset of a common category set. Especially, when there is no ambiguity about which entity in a visual scene is referenced[1], a visual recognition task is uniquely defined by specifying the range of a mapping. This intuition leads to a simple task modeling approach by treating a task as a set of answers. We exploit linguistic knowledge sources to extract the answer set. Specifically, we consider the following two sources: 1) a structured lexical database called Wordnet (Fellbaum, 1998) and 2) visual descriptions that are provided with a visual data.

Wordnet Wordnet (Fellbaum, 1998) is a lexical database represented with a directed acyclic graph of disambiguated word entities, called synsets. The graph represents a hierarchical structure of Wordnet, where the parents of a node correspond to hypernyms of the word in the child. We make a simple assumption that a set of words sharing common ancestors in the graph can construct an answer set in our concept hierarchy. A sample subgraph with extracted answer sets is illustrated in Figure 3. Given a list of answer sets extracted from Wordnet, a task specification $\mathbf{t}_s$ is a name of the answer set $\mathcal{W}_{\mathbf{t}_s}$ in the list. A task distribution is modeled by a distribution of answer set $\text{p}_{\mathcal{T}}(\mathbf{t}_s|\mathbf{a})$ given an answer $\mathbf{a}$, where a task specification $\mathbf{t}_s$ is independent of region annotation $\mathbf{r}$. We assign zero probability for answer sets that do not contain the target answer $\{\mathbf{t}_s|\mathbf{a} \notin \mathcal{W}_{\mathbf{t}_s}\}$ and uniform probability to all answer sets that contain the answer $\{\mathbf{t}_s|\mathbf{a} \in \mathcal{W}_{\mathbf{t}_s}\}$, where $\mathcal{W}_{\mathbf{t}_s}$ denotes an answer set named $\mathbf{t}_s$. Word embedding is used to encode $\mathbf{t}_s$ into a task specification vector $\tau_{\eta^s_{\text{pre}}}(\mathbf{t}_s)$.

[1] In the proposed VQA model, ambiguity of reference is usually resolved by an attention model.

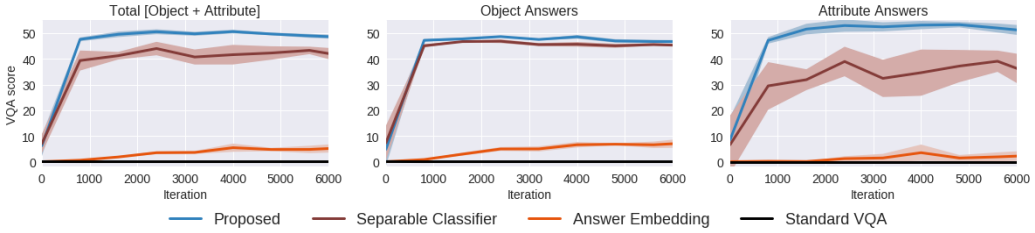


Figure 4: **Model comparisons.** Exploiting external data with unsupervised task discovery boosts performance of the proposed model and separable classifier significantly. However, separable classifier showed limited performance gain on attribute answers, which have significant variations depending on tasks.

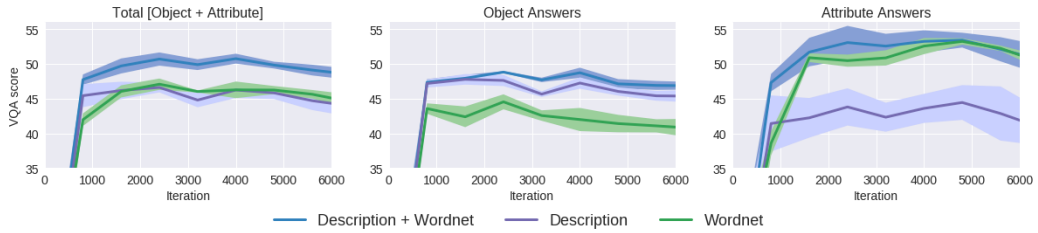


Figure 5: **Data comparisons.** Using visual description and Wordnet shows different generalization characteristics and combining them brings additional improvement.

Visual description We construct an answer set by selecting a word from a description and identifying possible alternatives of the selected one. Specifically, we employ a blanked description $\mathbf{t}_d = [w_1, ..., w_{T_d}]$ as a task specification, which is constructed by spotting an answer $\mathbf{a}$ from a description and replacing it to a special token <blank>. We define a task distribution $\mathrm{p}_{\mathcal{T}}(\mathbf{t}_d|\mathbf{a}, \mathbf{r})$ as a distribution of description $\mathrm{p}(\mathbf{d}|\mathbf{a}, \mathbf{r})$ because a blanked description $\mathbf{t}_d$ is deterministically constructed from a description $\mathbf{d}$ and an answer $\mathbf{a}$. Note that the distribution of description $\mathrm{p}(\mathbf{d}|\mathbf{a}, \mathbf{r})$ is determined by a dataset. Given $\mathbf{t}_d$, a task specification vector $\tau_{\eta^d_{\text{pre}}}(\mathbf{t}_d)$ is encoded by a gated recurrent unit (Chung et al., 2014).

5 Experiments

We evaluate how effectively the proposed framework leverages the external data without questions to answer out-of-vocabulary words in visual question answering. We compare the proposed method with baselines equipped with idea from zero-shot image classification (Frome et al., 2013) and novel object captioning (Anne Hendricks et al., 2016; Venugopalan et al., 2017), which are related to the proposed problem. We also analyze the impact of the external data used for pretraining and visualize the mapping between questions and task specifications learned by weakly supervised task regression.

5.1 Datasets

Pretraining We learn visual concepts about most frequently observed 3,000 objects and 1,000 attributes in Visual Genome dataset (Krishna et al., 2017). We construct external visual data with region bounding box annotations, which are provided with region descriptions. Then, we extract visual words—answer candidates—from region descriptions and construct visual data pairs $(\mathbf{r}, \mathbf{a})$. We also construct blanked description $(\mathbf{d})$ from the region description by replacing visual words with <blank>. Note that distribution $\mathrm{p}(\mathbf{d}|\mathbf{a}, \mathbf{r})$ is either zero or one depending on whether the description $\mathbf{d}$ is from the region $\mathbf{r}$ or not. We use 1,169,708 regions from 80,602 images for training data construction. To use Wordnet (Fellbaum, 1998), we map visual words to synset using synset annotations from visual genome dataset, and words that are not covered by the annotation are mapped using Textblob (Loria, 2018) python library.

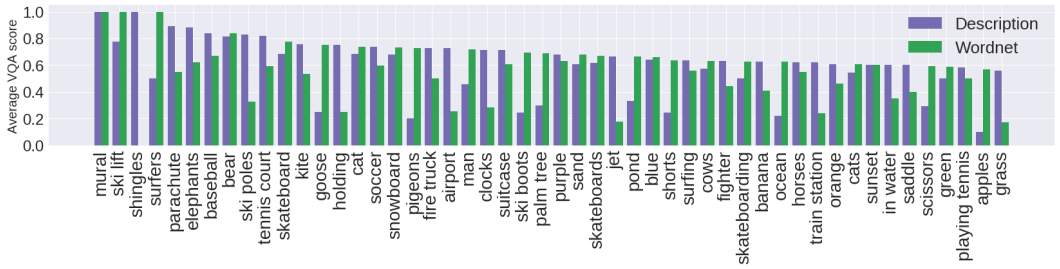


Figure 6: **Complementary characteristics of visual description and Wordnet.** Wordnet shows advantage on answers related to specific categorizations such as species of a bird (*e.g.* goose and pigeon) and visual description is more effective on answers about interactions (*e.g.* holding).

Dataset construction We repurpose VQA v2 dataset to construct a train/test split as illustrated in Figure 1. We use training and validation set of VQA v2. To ensure that every out-of-vocabulary answer appears during pretraining, we select out-of-vocabulary answers from the pretrained visual words. Among 3,813 visual words observed in VQA dataset, we randomly select 954 answers as out-of-vocabulary answers. Since we focus on transferability of visual words, answers about yes/no and numbers are not considered in our evaluation. Based on the selected out-of-vocabulary answers, we split questions into 462,788 training, 51,421 validation, 5,176 test-validation and 20,802 test splits. The training and validation splits do not contain out-of-vocabulary answers while test-validation and test splits consist of out-of-vocabulary answers only. We plan to make our splits publicly available. Evaluation is performed on the test split using the standard VQA protocol with 10 ground-truth answers annotated for each question (Antol et al., 2015). The VQA score is given by $100 \cdot \min\left(\frac{1}{3}\sum_{i=1}^{10} \mathbb{1}(\text{gt}_i, \hat{\mathbf{a}}), 1\right)$, where $\hat{\mathbf{a}}$ is a predicted answer, gt_i is the i-th ground truth answer and $\mathbb{1}(\cdot)$ is an indicator function.

5.2 Baselines

Since leveraging external visual data for visual question answering with out-of-vocabulary answers has hardly been explored, there is no proper evaluation benchmark and we employ the following baselines to compare with the proposed model:

- **Answer embedding** employs idea from zero-shot image classification (Frome et al., 2013) that learns mapping from visual features to pretrained answer embedding. We use GloVe (Pennington et al., 2014) to embed each answer.
- **Separable classifier** adopts idea from novel object captioning (Anne Hendricks et al., 2016; Venugopalan et al., 2017) that learns visual and language classifier separately and combines them by element-wise sum of logits for joint inference. This baseline and the proposed model are trained with the same data.

5.3 Results

Model comparisons Figure 4 illustrates model comparison results with standard VQA, answer embedding and separable classifier. For this experiment, we perform VQA adaptation with 4 different random seed and plot mean and standard deviation of VQA accuracy. The standard VQA model achieves 0 VQA score because there is no clue for inferring out-of-vocabulary answers. Answer embedding baseline generalizes slightly better by exploiting similarity of answer words in the embedding space. However, we observe very marginal performance improvement because mapping visual features onto answer embedding space is learned with insufficient information available only in the subset of answers. Using off-the-shelf visual data and task specifications from linguistic knowledge sources dramatically improves performance both for the separable classifier baseline and the proposed model. However, independent consideration of visual data and task specifications as in separable classifier has a critical limitation to model joint interaction between task specification and visual features. Especially, this baseline illustrates substantially lower performance on attribute answers, which have significant variations depending on tasks. Note that the bias in the VQA training set cannot be exploited in the proposed evaluation setting, as the evaluation is performed with out-of-distribution answers only.

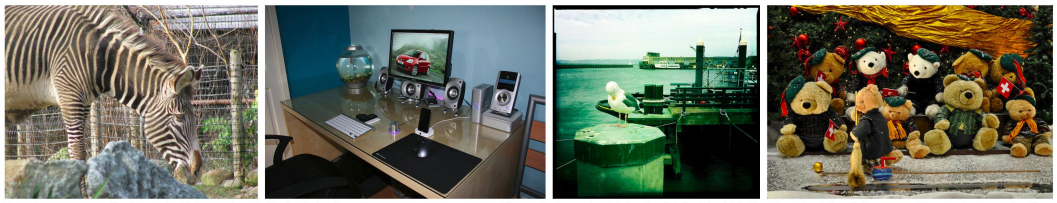


Figure 7: **Out-of-vocabulary answers with diverse types of concepts.** Green and red color denote correct and wrong answers respectively. Asterisk(*) denotes answers appearing in the training set. Answers without asterisks are out-of-vocabulary answers.

Table 1: **Learned mapping between a question and a task specification.** We retrieved questions for each answer sets based on the similarity score between task specification vectors. Results show that appropriate task specifications are regressed from each questions. No explicit supervision is used for mapping between questions and task specifications.

Answer set $\mathbf{t}_s$	Questions
organic_process.n.01	what are the giraffes doing? / what are the animals doing?
athletic_game.n.01	what type of sport ball is shown? / what type of sport are the men participating in?
furniture.n.01	what piece of furniture are the cats sitting on? / what furniture is the cat sitting on?
fruit.n.01	what type of fruit is the animal eating? / what type of fruit juice is on the counter?
time_period.n.01	what kind of season is it? / what type of season is it?
tool.n.01	what utensil is in the person 's hand? / what utensil is laying next to the bread?
hair.n.01	what hairstyle does the surfer have ? / what type of hairstyle does this man have ?

Data comparisons Figure 5 illustrates the effect of different linguistic sources on the proposed model. Visual description and Wordnet for the proposed model show complementary characteristics and additional improvement is achieved by their combination. To study the complementary characteristic of visual description and Wordnet, we visualize average VQA score for 50 answers in Figure 6. As shown in answers *goose* and *pigeon*, Wordnet has advantage on the questions about fine-grained classification such as species of a bird. This is because this categorization is explicitly modeled in the word hierarchy in Wordnet. On the other hand, visual description is more effective on the questions related to type of interactions, for example, *holding*.

Qualitative results Figure 7 shows prediction of proposed model and the baselines, where *Sep-cls* denotes separable classifier. The proposed model correctly predicts out-of-vocabulary answers for questions asking diverse visual concepts such as places, styles, species of bird and holiday names.

Weakly supervised task regression Given that answer sets extracted from Wordnet models diverse visual recognition tasks, matching these answer sets to relevant questions is critical for categorization of VQA data and model interpretation. As we learn VQA models by task regression, this matching can be performed by comparing the encoded task specification vector from a question $\tau_{\eta_{\text{vqa}}}(\mathbf{q})$ and the task specification vector of an answer set $\tau_{\eta_{\text{pre}}}(\mathbf{t}_s)$. For each $\tau_{\eta_{\text{pre}}}(\mathbf{t}_s)$, we sorted questions in a descending order of dot product similarity between $\tau_{\eta_{\text{pre}}}(\mathbf{t}_s)$ and $\tau_{\eta_{\text{vqa}}}(\mathbf{q})$. In the sorted question list, top 2 questions whose length fits in a table is visualized in Table 1. The proposed model is trained with Wordnet data for this experiment. Note that we perform the task regression without explicit supervision on matching between a question and a task specification.

6 Conclusion

We present a transfer learning approach for visual question answering (VQA) with out-of-vocabulary answers. We pretrain a task conditional visual classifier with off-the-shelf visual and linguistic data based on unsupervised task discovery. The pretrained task conditional visual classifier is transferred to VQA adaptively. The experimental results show that exploiting off-the-shelf visual and linguistic data boosts performance in the proposed setting and jointly training the task conditional visual classifier is important to model interaction between visual features and task specifications.

REFERENCES

Aishwarya Agrawal, Aniruddha Kembhavi, Dhruv Batra, and Devi Parikh. C-vqa: A compositional split of the visual question answering (vqa) v1. 0 dataset. *arXiv preprint arXiv:1704.08243*, 2017.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *CVPR*, 2018.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, Trevor Darrell, Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, et al. Deep compositional captioning: Describing novel object categories without paired training data. In *CVPR*, 2016.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. In *ICCV*, 2015.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pp. 722–735. Springer, 2007.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pp. 1247–1250. ACM, 2008.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Christiane Fellbaum. Wordnet: An electronic database, 1998.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*, 2016.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017.

Tanmay Gupta, Kevin Shih, Saurabh Singh, and Derek Hoiem. Aligned image-word representations improve inductive transfer across vision-language tasks. In *CVPR*, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017.

Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. In *ICLR*, 2016.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Skip-thought vectors. In *NIPS*, 2015.

Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2017.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, 2017.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

Steven Loria. TextBlob: Simplified Text Processing. `http://textblob.readthedocs.io/en/dev/`, 2018. [Online; accessed 3-May-2018].

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *CVPR*, 2018.

Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *NIPS*, 2014.

Hyeonwoo Noh, Paul Hongsuck Seo, and Bohyung Han. Image question answering using convolutional neural network with dynamic parameter prediction. In *CVPR*, 2016.

Dim P Papadopoulos, Jasper RR Uijlings, Frank Keller, and Vittorio Ferrari. We dont need no bounding-boxes: Training object class detectors using only human verification. In *CVPR*, 2016.

Dim P Papadopoulos, Jasper RR Uijlings, Frank Keller, and Vittorio Ferrari. Extreme clicking for efficient object annotation. In *ICCV*, 2017.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

Damien Teney and Anton van den Hengel. Zero-shot visual question answering. *arXiv preprint arXiv:1611.05546*, 2016.

Damien Teney, Peter Anderson, Xiaodong He, and Anton van den Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *CVPR*, 2018.

Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond Mooney, Trevor Darrell, and Kate Saenko. Captioning images with diverse objects. In *CVPR*, 2017.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. Pointer networks. In *NIPS*, 2015.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Fvqa: Fact-based visual question answering. *TPAMI*, 2017a.

Peng Wang, Qi Wu, Chunhua Shen, and Anton van den Hengel. The vqa-machine: Learning how to use existing vision algorithms to answer new questions. In *CVPR*, 2017b.

Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *CVPR*, 2016.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *CVPR*, 2016.

Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Incorporating copying mechanism in image captioning for learning novel objects. In *CVPR*, 2017.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *CVPR*, 2016.

A WEAKLY SUPERVISED TASK REGRESSION

This section describes how task regression is performed by Equation 2 in the main paper.

Let $\mathbf{q}$ be a question from a VQA dataset and $\tau^*_{\mathbf{q}}$ be a true task specification defined by a question. Our intuition is that a task is defined by a mapping from visual inputs to a set of answers, which is represented by a conditional distribution $\mathrm{p}_{\tau^*_{\mathbf{q}}}(\mathbf{a}|\mathbf{v})$. As the task is defined by a conditional distribution, the objective of a task regression becomes to find a task specification vector $\tau_{\eta_{\text{vqa}}}(\mathbf{q})$ that approximates this conditional distribution using a pretrained task conditional visual classifier $\mathrm{p}_\theta(\mathbf{a}|\mathbf{v}, \tau_{\eta_{\text{vqa}}}(\mathbf{q}))$. We can formulate this objective as a maximum log-likelihood, which is formally written as follows.

$$\mathbb{E}_{\mathrm{p}(\mathbf{v}|\mathbf{q})}\mathbb{E}_{\tau^*_{\mathbf{q}}(\mathbf{a}|\mathbf{v})}[\log\ \mathrm{p}_\theta(\mathbf{a}|\mathbf{v}, \tau_{\eta_{\text{vqa}}}(\mathbf{q}))], \tag{4}$$

where $\mathrm{p}(\mathbf{v}|\mathbf{q})$ is conditional distribution of visual features $\mathbf{v}$ given a question $\mathbf{q}$.

As we are interested in learning η_{vqa}, which is a parameter of a question encoder working over all question $\mathbf{q}$, we need to optimized the objective expected over the distribution of $\mathbf{q}$, which is formally written as follows.

$$\begin{aligned}
&\mathbb{E}_{\mathrm{p}(\mathbf{q})}\mathbb{E}_{\mathrm{p}(\mathbf{v}|\mathbf{q})}\mathbb{E}_{\tau^*_{\mathbf{q}}(\mathbf{a}|\mathbf{v})}[\log\ \mathrm{p}_\theta(\mathbf{a}|\mathbf{v}, \tau_{\eta_{\text{vqa}}}(\mathbf{q}))] \\
&\quad= \mathbb{E}_{\mathrm{p}(\mathbf{q})}\mathbb{E}_{\mathrm{p}(\mathbf{v}_{\phi_{\text{vqa}}}(\mathbf{I},\mathbf{q})|\mathbf{q})}\mathbb{E}_{\tau^*_{\mathbf{q}}(\mathbf{a}|\mathbf{v}_{\phi_{\text{vqa}}}(\mathbf{I},\mathbf{q}))}[\log\ \mathrm{p}_\theta(\mathbf{a}|\mathbf{v}_{\phi_{\text{vqa}}}(\mathbf{I}, \mathbf{q}), \tau_{\eta_{\text{vqa}}}(\mathbf{q}))] \\
&\quad= \mathbb{E}_{\mathrm{p}(\mathbf{q})}\mathbb{E}_{\mathrm{p}(\mathbf{I}|\mathbf{q})}\mathbb{E}_{\tau^*_{\mathbf{q}}(\mathbf{a}|\mathbf{I},\mathbf{q})}[\log\ \mathrm{p}_\theta(\mathbf{a}|\mathbf{v}_{\phi_{\text{vqa}}}(\mathbf{I}, \mathbf{q}), \tau_{\eta_{\text{vqa}}}(\mathbf{q}))] \\
&\quad= \mathbb{E}_{\mathrm{p}_{\text{vqa}}(\mathbf{a},\mathbf{I},\mathbf{q})}[\log\ \mathrm{p}_\theta(\mathbf{a}|\mathbf{v}_{\phi_{\text{vqa}}}(\mathbf{I}, \mathbf{q}), \tau_{\eta_{\text{vqa}}}(\mathbf{q}))],
\end{aligned} \tag{5}$$

where, $\mathrm{p}_{\text{vqa}}(\mathbf{a}, \mathbf{I}, \mathbf{q}) = \tau^*_{\mathbf{q}}(\mathbf{a}|\mathbf{I}, \mathbf{q})\mathrm{p}(\mathbf{I}|\mathbf{q})\mathrm{p}(\mathbf{q})$ and a visual feature $\mathbf{v}$ is altered to $\mathbf{v}_{\phi_{\text{vqa}}}(\mathbf{I}, \mathbf{q})$ because the visual feature is inferred by a visual encoder from an image $\mathbf{I}$. This derivation relates the task regression to standard VQA objective in Equation 1 in the main paper.

B COMBINING KNOWLEDGE LEARNED BY VQA

While we focus on learning visual concepts from external visual data, VQA dataset is still a valuable source of learning diverse knowledges. Especially, some answers in the VQA dataset are not visual words and require visual reasoning. For example, yes and no are one of the most frequent answers in the VQA dataset Antol et al. (2015) but it is not straightforward to learn these answers only with the external visual data. Therefore, we consider combining knowledge learned from VQA and from external visual data.

We construct a split of the VQA dataset consisting of 405,228 training, 37,031 validation, 43,171 test-validation and 172,681 test questions. The training and validation set does not contain any out-of-vocabulary answers and test-validation and test set contains out-of-vocabulary answers. Contrary to the split used in the main paper, this split also contains training answers in the test-validation and test set and the training answers includes logical answers, numbers and visual words. The list of out-of-vocabulary answers are identical to the main paper. Among 172,681 test questions, 103,013 questions could be correctly answered only with the training answers.

To combine knowledge from VQA and external visual data, we construct two task conditional visual classifier and fine-tune one classifier to learn newly appearing answers from VQA training set and fix one classifier to answer known visual answers including out-of-vocabulary answers. After training, we simply combine two logits by element-wise sum and pick answer with largest score for prediction.

The result is illustrated in Figure 8. We compare the proposed model with the standard VQA model and the answer embedding baseline. Each models are trained with 3 different random seeds and their mean and standard deviation are plotted. Overall, the proposed model performs the best. While the standard VQA model achieved the best performance for training answers, it cannot predict any out-of-vocabulary answers. The answer embedding baseline achieves some generalization to out-of-vocabulary answers, but constraints in the answer embedding degrades its performance on training answers.

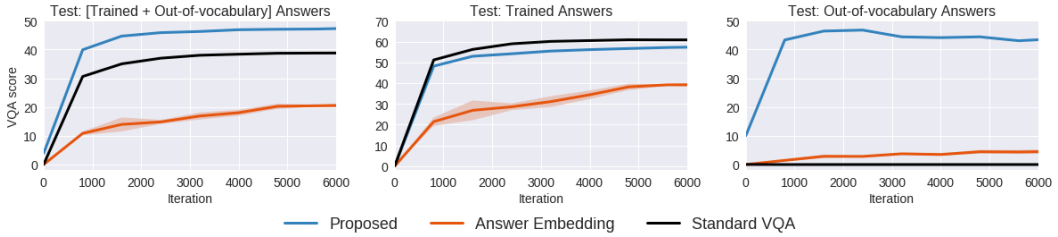


Figure 8: **Combining knowledge from VQA and external visual data.** Evaluation results on a test set containing both out-of-vocabulary answers and trained answers. The proposed model showed relatively lower performance on trained answers but significantly better performance on out-of-vocabulary answers. In total, the proposed model showed the best performance.

Table 2: **Learned mapping between a question and a task specification.** We retrieved questions for each answer sets based on the similarity score between task specification vectors. Results show that appropriate task specifications are regressed from each questions. No explicit supervision is used for mapping between questions and task specifications.

Answer set $\mathbf{t}_s$	Questions
animal.n.01	what type of animals are near the road? / what species of animals are these?
building_material.n.01	what kind of material is the flooring made from? / what is the type of flooring made of?
meat.n.01	what kind of meat is next to the broccoli? / what kind of meat is next to the veggies?
liquid.n.01	what kind of soda are the people drinking? / what type of soda are they drinking ?
plant.n.02	what kind of flower is in the tall vase? / what kind of plant leaves are on the plate?
plant_organ.n.01	what type of fruit is the animal eating? / what kind of fruit is the kid eating?
structure.n.01	what type of structure are they in? / what kind of building structure is she in?
furniture.n.01	what piece of furniture are the cats sitting on? / what furniture is the cat sitting on?
slope.n.01	is the man going uphill or downhill? / is the bus parked uphill or downhill?
color.n.01	what color is the trash bag? / what color is the garbage bag?
consumption.n.01	what are the cows doing? / what are the animals doing?
artifact.n.01	what object is the cat laying under? / what type of structure are they in?
fruit.n.01	what type of fruit is the animal eating? / what type of fruit juice is on the counter?
building.n.01	what kind of building structure is she in? / what type of building is he standing in?
hair.n.01	what hairstyle does the surfer have ? / what type of hairstyle does this man have ?
communication.n.02	what type of language is on the buildings? / what type of language is on the signs?
body_of_water.n.01	what type of body of water is the man on? / what type of body of water is in photo?
tool.n.01	what utensil is in the person 's hand? / what utensil is laying next to the bread?
time_period.n.01	what kind of season is it? / what type of season is it?
appliance.n.02	what kind of appliance is the cat standing in? / what appliance is she standing next to?
public_transport.n.01	what type of transportation is passing by? / what type of vehicle is shown in the sign?
fabric.n.01	what type of fabric are the bears made of? / what type of fabric is the chair made of?
tree.n.01	what kind of trees are under all that snow? / what type of trees are the tall ones?
beverage.n.01	what kind of soda is on the desk? / what kind of soda is in the bottle?
home_appliance.n.01	what kind of appliance is the cat standing in? / what appliance is she standing next to?
consumer_goods.n.01	what clothing item is this person wearing? / what clothing item is the girl wearing?
cutlery.n.02	what type of utensil is on the tray? / what utensil is under the fork?
edible_fruit.n.01	what type of fruit is the animal eating? / what kind of fruit is the purple fruit?
shape.n.02	what shape are most of the windows? / what shape is the tall structure to the right?
meal.n.01	what meal of the day are these designed for? / what meal are these typically eaten for?
sport.n.01	which sport are they doing? / what type of sport ball is shown?

C Additional Task Regression Results

Table 2 shows examples of task regressed questions corresponding to each answers sets, which are represented as the synset of common hypernyms. Similarity is computed by dot product between task specification vector of the answer set embedding $\tau_{\eta_{\text{pre}}}(\mathbf{t}_s)$ and the task specification vector regressed by a question $\tau_{\eta_{\text{vqa}}}(\mathbf{q})$. We manually select top ranked questions that fit within the table in terms of the number of characters.

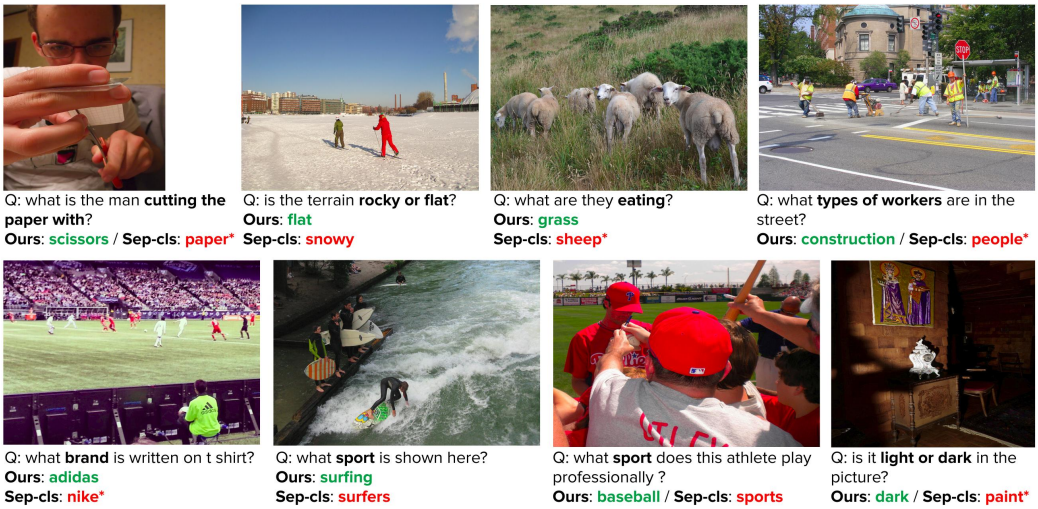
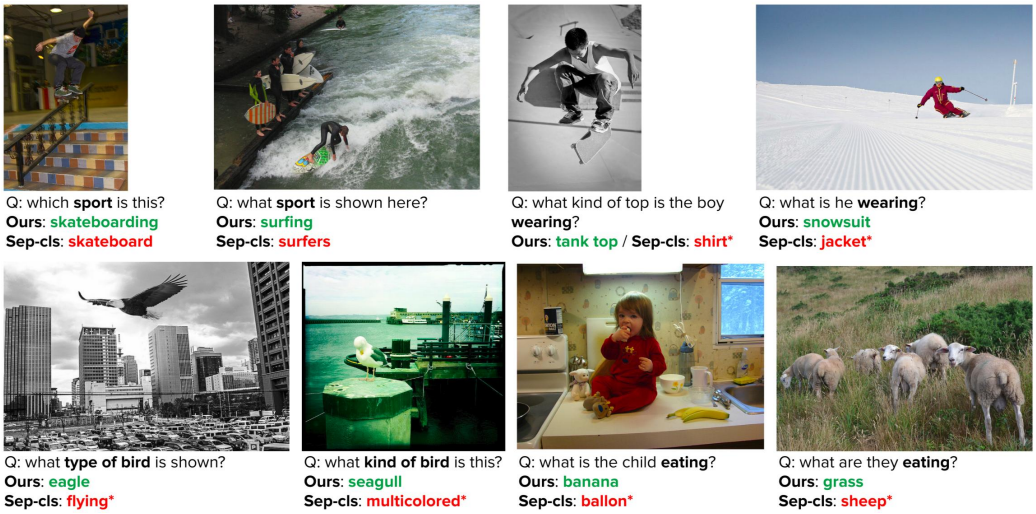


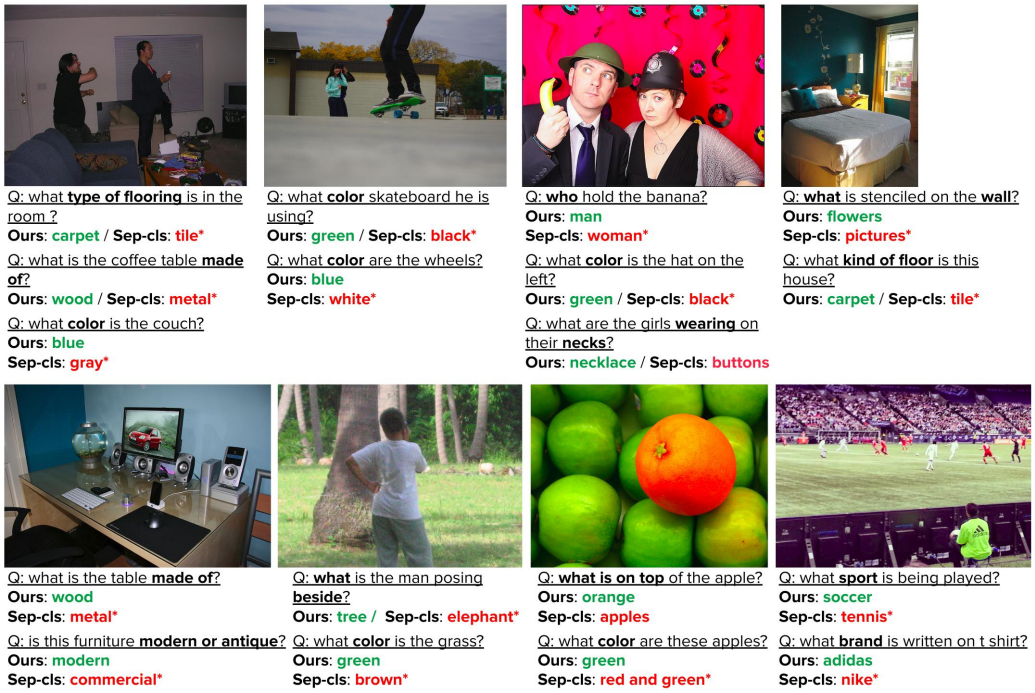
Figure 9: **Out-of-vocabulary answers with diverse types of concepts.** Green and red color denotes correct and wrong answers respectively. Asterisk(*) denotes answers appearing in the training set. Answers without asterisks are out-of-vocabulary answers. The proposed model correctly predict out-of-vocabulary answers for diverse visual recognition tasks.

D Additional Qualitative Results

To illustrate that the proposed model could answer diverse questions with out-of-vocabulary answers, we present additional qualitative results. Figure 9 illustrates that the proposed model correctly predict out-of-vocabulary answers to diverse visual recognition tasks. Figure 10 illustrates that the proposed model performs question and image dependent answering and could predict out-of-vocabulary answers depending on both image and question.



(a) **Same task with different out-of-vocabulary answers.** Pair of questions are asking similar visual recognition task. The proposed model correctly predict out-of-vocabulary answers depending on different images.



(b) **Same image with diverse tasks** This qualitative example visualizes diverse question answering for a single image. The proposed model correctly predicts out-of-vocabulary answers depending on the question.

Figure 10: **Image and question dependent answering.** Green and red color denote correct and wrong answers respectively. Asterisk(*) denotes answers appearing in the training set. Answers without asterisks are out-of-vocabulary answers.