Transfer Learning via Unsupervised Task Discovery for Visual Question Answering

Hyeonwoo Noh\textsuperscript{1,3} Taehoon Kim\textsuperscript{2*} Jonghwan Mun\textsuperscript{1,3} Bohyung Han\textsuperscript{3}
\textsuperscript{1}Computer Vision Lab. \textsuperscript{2}OpenAI \textsuperscript{3}Computer Vision Lab., ECE & ASRI
POSTECH, Korea USA Seoul National University, Korea
\textsuperscript{1}\{shgusdn, jonghwan.mun\}@postech.ac.kr \textsuperscript{2}carpedm20@gmail.com \textsuperscript{3}bhhan@snu.ac.kr

Abstract

We study how to leverage off-the-shelf visual and linguistic data to cope with out-of-vocabulary answers in visual question answering task. Existing large-scale visual datasets 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 can be captured and transferred to visual question answering models due to missing link between question dependent answering models and visual data without question. We tackle this problem in two steps: 1) learning a task conditional visual classifier, which is capable of solving diverse question-specific visual recognition tasks, based on unsupervised task discovery and 2) transferring 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 transfer a learned task conditional visual classifier as an answering unit in a visual question answering model. We empirically show that the proposed algorithm generalizes to out-of-vocabulary answers successfully using the knowledge transferred from the visual dataset.

1. Introduction

Human sees and understands a visual scene based on diverse visual concepts. For example, from a single image of a chair, human effortlessly recognizes diverse visual concepts such as its color, material, style, usage, and so on. Such diverse visual concepts may be associated with different questions in natural languages defining a recognition task for each of the visual concepts (e.g., what color is the chair?). Recently visual question answering (VQA) [5] 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 [3, 12, 43], 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 [1, 13] suffer from inherent bias, which hinders learning true visual concepts from the datasets.

On the contrary, human answers a question based on visual concepts learned from diverse sources such as books, pictures, videos, and personal experience that 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 image descriptions. Such information is already available in large-scale [9, 24, 22] and can scale further with reasonable cost [31, 32]. 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 a VQA problem with out-of-vocabulary answers, which is illustrated in Figure 1. External visual dataset provides 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 task is to handle out-of-vocabulary answers \( a \in \mathcal{A} - \mathcal{B} \) successfully by exploiting visual concepts learned from external visual dataset.

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 task conditional visual classifier, whose task is defined by a task feature. The classifier is used as an answering unit, where a task feature is inferred from a question. To train the task conditional visual classifier without

\*This work is performed while at Devsisters.
task annotations, we propose an unsupervised task discovery technique based on linguistic knowledge sources such as structured lexical databases, e.g., WordNet [10] and region descriptions [22]. We claim that the proposed transfer learning framework 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 show that the proposed approach handle out-of-vocabulary answers through knowledge transfer from external visual data.

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

The standard VQA evaluation assumes identically distributed train and test set [5, 29, 45]. As this evaluation setting turns out to be vulnerable to models exploiting biases in training set [13], several alternatives have been proposed. One approach is to reduce observed biases either by balancing answers for individual questions [13] or by providing different biases to train and test sets intentionally [1]. Another approach is to construct compositional generalization split [2, 18] 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 [2] or by constructing a synthetic dataset [18]. The problem setting studied in this paper is similar to [19, 36] 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 [15, 23] pre-trained on ImageNet [9] is a widely accepted standard for diverse VQA models [12, 43]. As an alternative, object detector [34] trained on the Visual Genome dataset [22] is employed to extract pretrained visual features [3]. Pretrained language models such as word embeddings [33] or sentence embeddings [21] are frequently used to initialize parameters of question encoders [12, 30, 35]. Exploiting information retrieval from knowledge base [6, 7] or external vision algorithms [40] to provide additional inputs to VQA models are investigated in [39, 40, 41]. Transfer between VQA datasets is studied in [16]. Sharing aligned image-word representations between VQA models and image classifiers has been proposed in [14] to exploit external visual data.

Transfer learning from external data to cope with out-of-vocabulary words has hardly been studied in VQA, but is actively investigated in novel object captioning [4, 28, 37, 44]. For example, [4] and [37] decompose image captioning task 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 [38] and learn to point an index of word candidates [44] or an associated region [28], where the word candidates are detected by a multi-label classifier [44] or an object detector [28] trained with external visual data. However, 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.

Our problem setting is closely related to zero-shot learning [11, 17, 25, 26, 42], where out-of-vocabulary answers are considered in classification. Zero-shot learning aims to recognize objects or classes that are unseen during training. As it aims generalization to completely unseen classes, any exposure to zero-shot classes during training is strictly prohibited [42]. On the contrary, our goal is to exploit class labels available in external dataset.

3. Transfer Learning Framework

The main goal of our work is to handle out-of-vocabulary answers in VQA by learning visual concepts from off-the-shelf visual dataset and transferring the concepts to VQA for answering the questions. Inspired by the fact that VQA can be thought of as a task conditional classification problem, where tasks are defined by questions, we introduce a task conditional visual classifier, which generates an answer from a visual input and a task specification, as a medium for learning and transferring visual concepts. Figure 2 illustrates the overall framework of the proposed approach. We pretrain the task conditional visual classifier using visual dataset without questions or task specifications via unsupervised task discovery, and adapt it to VQA models by transferring the learned parameters. We describe the task conditional visual classifier and how it is pretrained and transferred to VQA in the rest of this section.

3.1. Task conditional visual classifier

Task conditional visual classifier is a function taking a visual feature \(v \in \mathbb{R}^d\) and a task feature \(\tau \in \mathbb{R}^k\) and producing a probability distribution of answers or labels \(a \in [0, 1]^l\), where the terms answer and label are used interchangeably based on context hereafter. The classifier is formulated as a neural network with parameter \(\theta\) which models a conditional distribution \(p_\theta(a|v, \tau)\). Note that two inputs \(v\) and \(\tau\) are typically obtained by encoders \(v_{\phi_v}(\cdot)\) and \(\tau_{\eta_{\tau}}(\cdot)\).

In the proposed transfer learning scenario, a task conditional visual classifier is pretrained with off-the-shelf visual dataset, e.g., Visual Genome [22], and transferred to VQA. In the pretraining stage, the parameters of the classifier and two feature encoders \(\theta, \phi_{\text{pre}}\) and \(\eta_{\text{pre}}\) are jointly learned. This stage allows the task conditional visual classifier to handle diverse visual recognition tasks by learning the task feature \(\tau\). Transfer learning to VQA is achieved by reusing parameter \(\theta\) and adapting new encoders \(v_{\phi_v}(\cdot)\) and \(\tau_{\eta_{\tau}}(\cdot)\) to the learned task conditional visual classifier.

3.2. Pretraining

Learning the task conditional visual classifier is naturally formulated as the problem to maximize the following expected log likelihood:

\[
\theta^*, \phi^*, \eta^* = \arg\max_{\theta, \phi_{\text{pre}}, \eta_{\text{pre}}} \mathbb{E}_{D} \left[ \log p_{\theta}(a | v_{\phi_{\text{pre}}}(I, b), \tau_{\eta_{\text{pre}}}(t)) \right],
\]

where \(v_{\phi_{\text{pre}}}(I, b)\) is a visual feature based on an image \(I\) and a bounding box \(b\), and \(\tau_{\eta_{\text{pre}}}(t)\) is a task feature encoded from a task specification \(t\), \(a\) is an answer sampled from data distribution and it satisfies \(a \in A\). and \(\{\theta, \phi_{\text{pre}}, \eta_{\text{pre}}\}\) are model parameters. We obtain \(v_{\phi_{\text{pre}}}(I, b)\) using a learnable attention network parametrized by \(\phi_{\text{pre}}\) on top of off-the-shelf feature extractor [3], where the bounding box position \(b\) is used as a key to the attention. The optimization in Eq. (1) requires a joint distribution, \(p_{D}(a, I, b, t)\), which is not accessible in the external datasets in our setting due to missing task specifications \(t\). Section 4 describes how to model the joint distribution \(p_{D}(a, I, b, t)\) with the visual
3.3. Transfer learning for VQA

As illustrated in Figure 2, the proposed VQA model contains a task conditional visual classifier \( p_\theta(a|v, \tau) \). The pre-trained visual concepts are transferred to VQA by sharing the learned parameters \( \theta \). Then, learning a VQA model is now formulated as learning input representations \( v \) and \( \tau \) for \( p_\theta(a|v, \tau) \), which is given by

\[
\phi_{\text{vqa}}^*, \eta_{\text{vqa}}^* = \arg\max_{\phi_{\text{vqa}}, \eta_{\text{vqa}}} \mathbb{E}_{\eta_{\text{vqa}}} \left[ \log p_{\theta}(a|v_{\phi_{\text{vqa}}}(I, q), \tau_{\eta_{\text{vqa}}}(q)) \right],
\]

where \( v_{\phi_{\text{vqa}}}(I, q) \) is an encoded visual feature with an image \( I \) and a question \( q \) using an attention mechanism with parameter \( \phi_{\text{vqa}} \) and a off-the-shelf feature extractor [3]. A task feature \( \tau_{\eta_{\text{vqa}}}(q) \) encodes a question \( q \) using parameter \( \eta_{\text{vqa}} \). The joint distribution of a training dataset for VQA, \( p_{\text{vqa}}(a, I, q) \), is required for optimization, where answers from the distribution satisfy \( a \in A - B \). We learn \( \phi_{\text{vqa}} \) and \( \eta_{\text{vqa}} \) by maximizing the likelihood of the objective in Eq. (2) while the parameter for the pretrained task conditional visual classifier \( \theta \) remains fixed.

**Weakly supervised task regression** Utilizing a pre-trained task conditional visual classifier for the visual recognition specified by a question \( q \) requires to infer an optimal task feature \( \tau_q^* \). This requirement introduces a learning problem—task regression—that optimizes an encoder \( \tau_{\eta_{\text{vqa}}}(q) \) to predict \( \tau_q^* \) correctly. Because directly minimizing error \( E(\tau_q^*, \tau_{\eta_{\text{vqa}}}(q)) \) requires additional supervision about the tasks, we instead exploit VQA data as a source of weak supervision. We optimize indirect loss 

\[-E_{\tau_q}(a|v) \log p_\theta(a|v, \tau_{\eta_{\text{vqa}}}(q)),\]

which encourages the answer distributions conditioned on \( \tau_q^* \) and \( \tau_{\eta_{\text{vqa}}}(q) \) to be similar. By assuming that true task conditional answer distribution \( \tau_q(a|v) \) is implicitly modeled in VQA dataset, we employ Eq. (2) as the objective function for weakly supervised task regression.

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

**Matching visual features** To reuse pretrained visual classifier in VQA without fine-tuning, semantics of visual features \( v \) should not change by learning with VQA dataset. This is fulfilled in recent approaches for VQA models that do not fine-tune pretrained visual feature extractors and focus on learning attention mechanism [20] over the extracted feature map. In our setting, we simply use the identical visual feature extractor [3] for both pretraining and VQA.

4. Unsupervised Task Discovery

Learning a task conditional visual classifier with the off-the-shelf visual dataset [22] is not straightforward due to missing annotation for task specifications, which is necessary to learn an encoder for a task specification vector \( \tau \). To address this issue, we propose unsupervised task discovery which samples a task specification \( t \) from a task distribution modeled by exploiting linguistic knowledge sources.

**4.1. Leveraging linguistic knowledge sources** A visual recognition task given by a question typically defines a mapping from visual inputs to a set of possible visual concepts, that is, a word group. For example, a question of “what is the woman holding?” defines a visual recognition task finding holdable objects in an image, which is a classification over a word group \{ball, racket, cup, ...\}. This intuition leads to a simple approach for modeling distribution of task description \( t \) by treating a task as a word group (i.e. holdable objects). The main reason to use linguistic knowledge sources for the unsupervised task discovery is that the word groups are often accessible in the linguistic knowledge sources. We consider two linguistic knowledge sources: 1) visual descriptions provided with a visual data and 2) a structured lexical database called WordNet [10]. Figure 3 illustrates the overview of our approach.

![Figure 3](image-url)
The graph represents a hierarchical structure of words, where the parents of a node correspond to hypernyms of the word in the child. In WordNet, we define a task specification \( t_w \) as a synset for a node that is a common ancestor of multiple words because a set of words sharing a common ancestor constructs a word group and the word group may also define a visual recognition task.

A procedure for sampling a task specification \( t_w \) based on WordNet and a visual data \( (a, b, d) \) is illustrated in Figure 3b. The main idea for the procedure is to model a task distribution conditioned on an answer \( p(t_w|a) \) as a uniform distribution over possible word groups that the answer belongs to, where a task specification \( t_w \) is a common ancestor of words in an word group. Modeling the distribution \( p(t_w|a) \) requires two stages: 1) constructing a word group table, which maps a task specification to a word group and 2) constructing an inverted word group table, which maps an answer word to a set of task specifications. The inverted word group table is used to retrieve a set of possible task specifications for an answer \( a \) and the distribution \( p(t_w|a) \) is the uniform distribution over task specifications in the set. Given the distribution \( p(t_w|a) \), the joint distribution, \( p_D(a, b, t_w) = \int p(a, b, t_w, d) dd \), is given by

\[
p(a, b, t_w, d) = p(t_w|a)p_V(a, b, d)
\]

where \( p(t_w|a) = \delta(t_w, \rho(a)) \) is a delta function that returns 1 if two inputs are identical and 0 otherwise. As illustrated in Figure 3a, we sample data required for pretraining \( (a, b, t_w) \) by first sampling a visual data \( (a, b, d) \) and then sampling a task specification \( t_w \) from \( p(t_w|a) \). This procedure results in sampling description \( d \) as well, but we do not care about it when we pretrain the task conditional visual classifier. For pretraining, we encode \( t_w \) into a task feature \( \tau_{t_w}(t_w) \) based on a gated recurrent unit \( [8] \) because \( t_w \) is a sequence of words.

The main reason to use a blanked description for a task specification is that it is effective to define a set of candidate words. For example, a blanked description “a man is holding _____” restricts candidate words for the blank to a set of objects that can be held. Therefore, a blanked description can be used to determine a word group implicitly that represents a visual recognition task.

### 4.3. WordNet

WordNet [10] is a lexical database represented with a directed acyclic graph of disambiguated word entities, called synsets. A sample subgraph of WordNet is illustrated in Figure 4 (left). The graph represents a hierarchical structure of words, where the parents of a node correspond to synsets. A sample subgraph of WordNet is illustrated in Figure 4 (left). The graph represents a hierarchical structure of words, where the parents of a node correspond to synsets. A sample subgraph of WordNet is illustrated in Figure 4 (left). The graph represents a hierarchical structure of words, where the parents of a node correspond to synsets.

<table>
<thead>
<tr>
<th>task ( t_w )</th>
<th>word groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>freezer, hair dryer, refrigerator, fridge, oven, dishwasher, washing machine, ...</td>
<td></td>
</tr>
<tr>
<td>manhole, rear window, exit, nozzle, car window, coin slot, spout, vent, ...</td>
<td></td>
</tr>
<tr>
<td>potatoes, french fries, chicken, melted cheese, tomatoes, sausage, vegetable, ...</td>
<td></td>
</tr>
<tr>
<td>twisted, ruffled, curving, curved, wheeled, rolled, coiled, turned, rippled, ...</td>
<td></td>
</tr>
<tr>
<td>fishing, skateboarding, sailing, playing baseball, surfing, traveling, driving, ...</td>
<td></td>
</tr>
<tr>
<td>middle, in corner, center, parking space, playground, landscape, neighborhood, ...</td>
<td></td>
</tr>
<tr>
<td>yellow, pink, red, beige, royal blue, amber, aqua, dark red, olive green, teal, ...</td>
<td></td>
</tr>
</tbody>
</table>
Figure 5. **Model comparisons.** Exploiting external data with the unsupervised task discovery boosts performance of the proposed model and separable classifier significantly while the separable classifier shows limited gains on attribute answers with large variations.

Figure 6. **Data comparisons.** Using visual description and WordNet shows different generalization characteristics and combining them brings additional improvement.

Figure 7. **Complementary characteristics of data.** Visual description and WordNet show complementary characteristics in terms of VQA score for different answers.

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 the baselines equipped with idea for zero-shot image classification [11] and novel object captioning [4, 37], 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. Notably, the evaluation setting is not comparable to that from zero-shot learning literature [42], as our objective is not generalization to completely unseen classes, but exploiting class labels exposed in the external data. We open-sourced all the codes and datasets used in the experiments to facilitate reproducing the result in this paper.

### 5.1. Datasets

**Pretraining** We learn visual concepts of most frequently observed 3,000 objects and 1,000 attributes in the Visual Genome dataset [22]. To pretrain the task conditional visual classifier, we construct external visual data with bounding box annotations, which are provided with region descriptions. Then, visual words (answers) are extracted from region descriptions to construct visual data quadruples \((a, I, b, d)\). We use 1,169,708 regions from 80,602 images to construct the training data. To use WordNet [10], we map visual words to synsets using synset annotations from Visual Genome dataset, and the words that are not covered by the annotations are mapped to synsets using TextBlob [27].

**Dataset construction** We repurpose VQA v2 dataset to construct a training/test split as illustrated in Figure 1. We use training and validation set of VQA v2. To ensure that

1. https://github.com/HyeonwooNoh/vqa_task_discovery
Figure 8. Out-of-vocabulary answers with diverse types of concepts. Green color denotes correct answers. All predicted answers are from out-of-vocabulary answers. The proposed model successfully predicts diverse out-of-vocabulary answers depending on questions.

Table 1. Result of weakly supervised task discovery. We retrieve questions for each task specification based on the similarity score with task features. Results show that appropriate task specifications are regressed from each question. Note that no explicit supervision is used for learning the mapping between questions and task specifications.

<table>
<thead>
<tr>
<th>Task specification $t_w$</th>
<th>Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>organic_process.n.01</td>
<td>what are the giraffes doing? / what are the animals doing? / what are the giraffes doing in this picture?</td>
</tr>
<tr>
<td>athletic_game.n.01</td>
<td>what type of sport ball is shown? / what type of sport are the men participating in?</td>
</tr>
<tr>
<td>furniture.n.01</td>
<td>what piece of furniture are the cats sitting on? / what furniture is the cat sitting on?</td>
</tr>
<tr>
<td>fruit.n.01</td>
<td>what type of fruit is the animal eating? / what type of fruit juice is on the counter?</td>
</tr>
<tr>
<td>time_period.n.01</td>
<td>what kind of season is it? / what type of season is it? / which season is it?</td>
</tr>
<tr>
<td>tool.n.01</td>
<td>what utensil is in the person’s hand? / what utensil is laying next to the bread?</td>
</tr>
<tr>
<td>hair.n.01</td>
<td>what hairstyle does the surfer have? / what type of hairstyle does this man have?</td>
</tr>
</tbody>
</table>

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 our algorithm: 1) answer embedding, which employs idea from zero-shot image classification [11] that learns mapping from visual features to pretrained answer embedding, where we use GloVe [33] to embed each answer, and 2) separable classifier, which adopts idea from novel object captioning [4, 37] that learns visual and language classifier separately and combines them by element-wise sum of logits for joint inference. Note that the separable classifier and our proposed model are trained with the same data.

5.3. Results

Model comparisons Figure 5 illustrates model comparison results with the baselines. For this experiment, we perform VQA adaptation with 6 different random seeds and plot their mean and standard deviation. The standard VQA model fails to predict any out-of-vocabulary answers (i.e., 0 VQA score) because there is no clue for inferring out-of-vocabulary answers. Answer embedding baseline generalizes slightly better by exploiting the similarity of answer words in the embedding space, but the improvement is marginal. Using off-the-shelf visual data and task specifications from linguistic knowledge sources dramatically improves performance both on the separable classifier and the proposed model. However, independent consideration of visual data and task specifications in separable classifier has a critical limitation of modeling joint interaction between task specifications and visual features. Especially, this baseline shows 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.

Data comparisons Figure 6 illustrates the effect of different linguistic sources in our algorithm. Two models learned using visual description and WordNet, respectively, have complementary characteristics and additional improvement is achieved by exploiting both data. More detailed comple-
5.4. 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 are not visual words and require visual reasoning. For example, yes and no are one of the most frequent answers in the VQA dataset [5] but it is not straightforward to learn these answers only with the external visual data. Therefore, we consider combining knowledge learned from VQA dataset and from external visual data.

We construct a split of the VQA dataset consisting of 405,228 training, 37,031 validation, and 172,681 test questions. The training and validation set do not contain any out-of-vocabulary answers and test set contains out-of-vocabulary answers. However, contrary to the main experiment, the test set also contains training answers which include logical answers, numbers and visual words. The list of out-of-vocabulary answers are identical to that of the main experiment. Among 172,681 test questions, 103,013 questions can be answered only with the training answers.

To combine knowledges from VQA dataset and external visual data, we learn a VQA model with two task conditional visual classifiers; we fine-tune one classifier for adapting answers requiring visual reasoning (i.e., numbers and yes/no) and fix the other classifier for visual answers including out-of-vocabulary answers. After training the VQA model, we combine two logits by element-wise sum and pick the answer with the highest score in the inference.

The results are presented in Figure 9. Models in each method are trained with 6 different random seeds and their mean and standard deviation are plotted. Overall, the proposed model performs the best. While the standard VQA model achieves the best performance for training answers, it fails to predict any out-of-vocabulary answers. The answer embedding baseline somewhat generalizes to out-of-vocabulary answers, but constraints in the answer embedding degrade its performance on answers in training set.

6. Conclusion

We present a transfer learning approach for visual question answering 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 external visual and linguistic data boosts performance in the proposed setting and training with unsupervised task discovery is important to model interaction between visual features and task specifications.

Acknowledgments This research was partly supported by Kakao and Kakao Brain and Korean ICT R&D program of the MSIP/IITP grant [2016-0-00563, 2017-0-01778].
References


[31] 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. 1


[43] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked Attention Networks for Image Question Answering. In *CVPR*, 2016. 1, 2


8394