Abstract

Collecting fully annotated image datasets is challenging and expensive. Many types of weak supervision have been explored: weak manual annotations, web search results, temporal continuity, ambient sound and others. We focus on one particular unexplored mode: visual questions that are asked about images. The key observation that inspires our work is that the question itself provides useful information about the image (even without the answer being available). For instance, the question “what is the breed of the dog?” informs the AI that the animal in the scene is a dog and that there is only one dog present. We make three contributions: (1) providing an extensive qualitative and quantitative analysis of the information contained in human visual questions, (2) proposing two simple but surprisingly effective modifications to the standard visual question answering models that allow them to make use of weak supervision in the form of unanswered questions associated with images and (3) demonstrating that a simple data augmentation strategy inspired by our insights results in a 7.1% improvement on the standard VQA benchmark.

1. Introduction

Supervised learning has shown great promise in developing visual AI. However, collecting manually annotated visual datasets is both challenging and expensive [36, 14, 28]. Using cheaper and weaker supervision is a growing research direction [46, 42, 8, 12, 35, 7, 34, 4]. As AI is increasingly integrated into our daily lives, computer vision systems will have access to increasingly diverse sources of information by constantly observing human-human, human-object, human-environment and human-AI interactions. Efficiently utilizing all this information will become increasingly critical as we aim to develop large-scale, accurate and adaptable visual systems.

Visual Question Answering (VQA) has become a new mode of interaction between humans and AI [3]. Presently, VQA is mostly used as means for evaluating visual reasoning capabilities of computers. However, going forward this is likely to become a natural human-AI interaction paradigm. It will become commonplace for humans to ask computers visual questions, such as, “Where did I leave my keys?”, “What breed of dog is this?”, “Have I met this person before?” or “Why is she doing that?” Instead of viewing this as single-sided interaction with humans soliciting information from AI systems, we consider how visual questions themselves can serve as a form of supervision to improve computer vision systems (Figure 1).

In contrast to existing works that focus on improving AI’s VQA capabilities [16, 26, 33], we strive to understand how much information is contained within the question itself, even when the answer is not provided (as would be the case of human-AI conversations). E.g., the question “What breed of dog is this?” provides information that the animal in the scene is a dog and suggests that there is a single dog present. The question “Why is he doing that?” suggests that the depicted behavior is unusual or unexpected. This type of free, natural and open-ended supervision can pave the way to developing richer cognitive AI.

We set out to investigate this hypothesis that human questions can be effectively used to improve computer vision capabilities. We begin by providing extensive qualitative and quantitative analysis of the information contained in a visual question using the large-scale VQA dataset [3]. We propose two simple but surprisingly effective modifications to the iBOWIMG [48] VQA model that allows it to make use of weak supervision in the form of images associated with unanswered questions. This proves our hypothesis that unanswered questions can be effectively used as a form of visual supervision.

Inspired by the insights from our initial experiments,
we then propose a simple data augmentation strategy. The key idea is that instead of using the image-question-answer triplet as a training exemplar, we generate $2^n$ training exemplars incorporating all possible subsets of the $n$ questions associated with the image. This strategy yields a 7.1% improvement in accuracy on the standard VQA benchmark, which confirms that our analysis has important implications not just in the future of close AI-human interactions but for the immediately relevant benchmarks.

Our code, models and additional details are available at http://sidgan.me/whats_in_a_question/.

2. Related work

**Vision and language:** Computer vision models are generally trained to recognize a fixed vocabulary of visual concepts [14, 28, 36]. But recently, there has been a trend towards more descriptive open-world image understanding. Efforts have included works on image [41, 15, 9, 40, 23, 25] and video [13] captioning, image segmentation from natural language expressions [19], aligning videos and books [49], zero-shot recognition from natural text [5, 18], learning object models from noisy open-world human labels [32] and other methods. While visual questions are certainly not meant to provide a complete description of the image, they still contain some open-world information about the scene encoded in natural text. In this work we take the initial steps towards extracting and harnessing this information.

**Visual question answering:** The literature on building visual question answering systems [6, 47, 45, 22, 18, 24, 31, 2, 38, 16, 33, 20, 43, 44, 48, 37, 26] is far too extensive to be covered in detail here. We do some analysis by building off of the iBOWIMG model of Zhou et al. [48]. But, much of our investigation is orthogonal to the visual question answering pipeline. Our work on understanding the information embedded within a question is more similar to works such as Lin et al. [29] on utilizing VQA knowledge to improve image captioning or Goyal et al. [17] on analyzing the relative informativeness of different words within a question. However, in contrast to those approaches, we focus on the knowledge that can be extracted from the question alone and not the question-answer combination.

**Incidental supervision:** As we move towards large-scale open-world visual understanding, collecting manually annotated datasets for every task and concept is quickly becoming infeasible. Developing ways of using natural and cost-effective forms of supervision is a growing research direction: weak manual annotations [46, 42], web search-based supervision [8, 12], or extra modalities like temporal continuity [35], depth [7], ambient sound [34] or GPS signal [4]. Along similar lines, we investigate whether visual questions associated with images provide sufficient supervision to train computer vision models. We argue that increasing integration of AI into everyday human environments will organically generate a large set of image-question pairs that can be used to improve visual AI systems.

3. Inspiration: What information do unanswered questions contain?

We begin with qualitative and quantitative analysis of the information visual questions may contain. We consider a setting where we have an image and a question (or a set of questions) associated with it, but with no corresponding answer. We examine the information content of the questions from two perspectives: (1) whether these questions can provide a good image description and (2) whether we can learn what objects are present in the image, given these questions. The insights from our analysis provide inspiration for the method described in Section 4 for utilizing the unanswered questions in learning vision models.

**Setup:** We detail the setup for analysis here. We use the COCO dataset with 82,783 training and 40,504 validation images [28]. Three types of annotations are associated with the dataset: (1) visual questions, where each image is associated with three human-generated questions about the visual scene [3]. (2) image captions, where every image is associated with five human-generated natural language descriptions and (3) image classification labels, where every image is annotated with the presence or absence of 80 target object classes. In this section we don’t make use of the answers to the visual questions.

3.1. Image description

Image captions are a natural open-world way to describe an image. [3] qualitatively notes the difference in information between image captions and visual questions: questions tend to provide specific information regarding one object within the image, while captions naturally tend to be a richer source of information. However, when a human looks at a scene, it is rare that she will be compelled to provide a caption (except when posting the image on social media). In contrast, she may feel compelled to ask a question, such as, “Is this rice noodle soup?” or “Are the flowers real or artificial?” The fact that she asks these questions provides some information about the scene contents. We begin by analyzing whether the visual questions contain enough information to provide an accurate description of the image.

**Quantitative results:** We evaluate using visual questions as image captions in Table 1 using two standard captioning metrics: METEOR [11] and SPICE [1]. We first consider three baselines that don’t use image information but generate a caption purely based on the visual questions that have been asked: (1) **One Q:** using one of the visual questions directly as a caption, (2) **Three Qs:** using all three visual questions concatenated together as a caption and (3) **Seq2Seq** [10]: a model trained on the COCO training set.
that takes an input of three visual questions and learns to output an image caption based on the information contained in the questions. Three Qs outperforms One Q (SPICE score of 0.115 vs 0.058), indicating that different questions provide complementary information about the image content.\(^1\)

Training the Seq2Seq model to generate more semantically meaningful captions from the three questions provides an improvement to the SPICE score from 0.115 to 0.140.\(^2\)

We additionally investigate whether visual questions can provide complementary information to what is contained in the image features. We use a computer vision model NeuralTalk2 (NT)\(^{[25]}\) that takes in an image and outputs an image caption. Directly concatenating this image-based caption with the caption generated from questions (NT + Seq2Seq) improves the SPICE score from 0.194 to 0.256, indicating that the signal from visual questions may be complementary to the information in the image.

**Qualitative results:** Finally, we qualitatively show some results of captions generated from visual questions. Figure 2 shows some results of applying the Seq2Seq model on the validation set to convert the 3 visual questions associated with the image to a single image caption. The results demonstrate that the visual questions can provide detailed information about the image content. The generated captions contain object category, human actions, color and affordance information indicating that this information can be readily extracted from the questions.

### 3.2. Object Classification

Besides image description, another source of information that the visual questions can provide is the object classes that are present in the image. Some examples are shown in Table 2: e.g., asking “what color is the bus?” indicates the presence of a bus in the image.

**Algorithm:** To quantify how often this occurs, we extract object labels for the 80 COCO classes from visual questions. There are 64 question types in COCO: “how many,” “is there,” “what color is”, etc. For each type, we manually determine if questions of this type imply the presence of objects. For example, questions of the type “how many” do imply the presence of objects: “how many different flowers are on the table?” implies the presence of flowers and a table. In contrast, questions of the type “is there,” such as “is there a zebra in the photo?”, do not imply the presence of any object. For each question that implies the presence of the object, we extract which of the 80 COCO classes (if any) the question refers to. We use NLTK\(^{[30]}\) to disambiguate tenses and synonyms as well as pattern.en\(^3\) for singular-plurals. For two-word categories such as “teddy bear” we use n-gram overlap.

**Results:** We compare the resulting object class vectors with ground truth annotations of object presence in the image. Our conversion algorithm achieves mean per-class recall of 29.3% and precision of 82.4%, indicating that while the three visual questions do not refer to all objects in the image, they nevertheless capture more than a quarter of the

---

\(^1\)The SPICE metric\(^1\) considers both precision and recall of a caption, enabling a fair comparison between captions of different lengths.

\(^2\)Our analysis bears some similarity to the work of Lin et al.\(^{[29]}\) on re-ranking image captions using VQA; however, we don’t use the answers or the image and evaluate captions generated solely based on the questions.

\(^3\)http://www.clips.ua.ac.be/pages/pattern-en
common objects with a few false positives.

Figure 3 shows the per-class recall and precision as a function of the average size of this class in an image. As expected, objects which are larger tend to be asked about more frequently. For example, “baseball glove” occupies only 0.8% of the image on average and has a near-zero recall of 0.7%, indicating that it is never asked about (or we are not able to parse it out with our algorithm). In contrast, “train” occupies 25.8% of the image area and has a recall of 79.5%, indicating that if a train appears in the image it is almost always asked about. A notable exception is “dining table,” which occupies 31.7% of the image area on average but has a recall of only 16.3% since it is rarely a target object of interest. Overall across all classes, the objects we do detect occupy 18.2% of the image area on average, whereas the objects that we fail to detect occupy only 7.1%.

Combination with vision models: We additionally verify that knowing the questions provides extra information about the image beyond what can currently be extracted by modern computer vision models. We finetune an ILSVRC2012-pretrained GoogLeNet model \[39, 36\] on the training set of COCO to recognize the 80 target object classes. It achieves image classification mAP of 53.1% on the validation set. We then combine the 80-dimensional classifier prediction vector \(x_o\) with our object class vector \(x_c\) extracted from the three visual questions using \(\max(x_o, x_c)\). This significantly increases image classification accuracy to 67.2% mAP.

Discussion: We showed that visual questions, even without answers, provide informative image descriptions and object classification information. In addition, we briefly note that visual questions can also provide additional latent information as illustrated in Figure 4. We make no attempt to quantify here but note that this information may also potentially be extracted and exploited in the future AI systems.

4. Method: Effectively utilizing information from unanswered visual questions

Armed with the conclusion that visual questions themselves provide important and useful information about the image content, we now set out to investigate how these questions can be used to aid the development of improved computer vision models. We focus on the VQA task and investigate how even unanswered questions can be effectively utilized to improve VQA capabilities. Since our proposed formulation is very simple, the empirical benefits demonstrated in Section 5 are even more striking.

Standard VQA systems [6, 47, 45, 22, 18, 24, 31, 2, 38, 16, 33, 20, 43, 44, 48, 37] take the image and its target question, but also to a set of other questions to quantify here but note that this information may also potentially be extracted and exploited in the future AI systems.

4.1. Model

To investigate, we build upon the iBOWIMG model [48]. This model is perfect for our investigation as it is very
The training examples for iBOWIMG are of the form:

\[(x, q_{ij}, a_{ij}) \forall i, j\]  

(1)

In contrast, the training examples for iBOWIMG-2x are:

\[(x, q_{ij}, E, a_{ij}) \forall i, j, E \subseteq \mathcal{P}(Q_{i}^{all})\]  

(2)

where \(\mathcal{P}\) denotes the powerset of \(Q_{i}^{all}\) and defines the extra information provided to the model in the form of additional questions asked about the same image.

For example, consider an image \(x\) with a question \(q\), a corresponding answer \(a\) and two additional unanswered questions \(q'_{1}\) and \(q'_{2}\). For iBOWIMG, the single training example corresponding to this image would be \((x, q, a)\). For iBOWIMG-2x there would be eight training examples, with \(E = \emptyset, q, q'_{1}, q'_{2}, \{q, q'_{1}\}, \{q, q'_{2}\}, \{q'_{1}, q'_{2}\}\) or \([q, q'_{1}, q'_{2}]\) making use of the extra information that is available about this image during training in the form of other asked questions.\(^5\)

The target label for all these exemplars is the answer \(a_{ij}\).

After the new exemplars are generated, the model is trained using stochastic gradient descent exactly as iBOWIMG.

### 4.3. Unanswered questions on novel images

One disadvantage of the method described so far is that it can only incorporate information from extra questions on images that have at least one answered question provided. However, it may be the case that we have access to a large collection of images with only unanswered questions associated with them: e.g. A dataset of image-question-pairs without their associated ground truth can naturally emerge from a deployed VQA system that is interacting with users in the real world. Motivated by the findings of Section 3.2, we use \(Q_{i}^{all}\) to learn an image representation that may be better suited for the VQA task. Instead of using an ILSVRC-trained visual model, we use a visual model trained to recognize the words that appear within the questions. Intuitively, ILSVRC-trained models may not reflect the full spectrum of visual concepts or diverse visual scenes. This new image model can be incorporated into iBOWIMG-2x (or even iBOWIMG) as a better image representation.

### 4.4. Testing

The iBOWIMG-2x model can be evaluated in one of two ways. During test time for a standard VQA formulation, the model only has access to a novel image \(x\) and a single target question \(q\). In this case, we can simply pass a zero-initialized vector for the extra features, reducing iBOWIMG-2x back to iBOWIMG but trained differently. However, iBOWIMG-2x allows additional flexibility

\(^4\)Our model bears some similarity to that of [21] which explores a different setting, where they double the dimensionality of the bag-of-words textual representation. However, they concatenate the question, image and answer features to predict the correctness of such image-question-answer triplets. In contrast, our feature vector utilizes the image features, target question and other questions about the image.

\(^5\)While this model is formulated to make use of extra unanswered questions, an additional benefit is that it can be considered a form of data augmentation. For example, if 3 answered questions are available for this image, the iBOWIMG would have 3 training examples while iBOWIMG-2x would have 24 training examples.

---

**Figure 5:** Framework of the iBOWIMG-2x model. The representation consists of three parts: (1) visual image features, (2) text embedding of the target question, and (3) text embedding of the other questions concatenated together. This representation is passed through a learned fully connected layer to predict the answer to the target question.

The image and target question features are concatenated and sent through a softmax layer to predict the correctness of such image-question-answer triplets. In contrast, our feature vector utilizes the image features, target question and other questions about the image.
by utilizing unanswered questions even at test time. For example, when the test image is provided with several target questions, they can further help interpret the image: e.g., test questions “Who is to the left of the dog?” and “What is to the right of the person?” provide complementary information that might help answer both questions better.

5. Experiments

We now empirically verify our intuition that even unanswered questions can significantly improve the accuracy of VQA systems. In particular, we evaluate our proposed iBOWIMG-2x model trained on subsets of the COCO [28] dataset corresponding to two different settings: (1) where every image has at least one answered question and optional unanswered questions associated with it in Section 5.1, and (2) where some images have only unanswered questions associated with it in Section 5.2. We convincingly demonstrate that including extra questions significantly improves VQA accuracy. To conclude, we apply our insights to the standard VQA benchmark in Section 5.3.

Setup: We use the COCO dataset with 82,783 training and 40,504 validation images. Each image is associated with three questions and their corresponding answers, although we sometimes use only a subset of those in our experiments (details below). We evaluate the model on the multiple-choice VQA task. We normalize the visual features and the two textual features independently to have $L2$ norm of 1. We build upon the code released by [48].

5.1. Unanswered questions on training images

Dataset: Consider the setting where we have access to a set of training images, each with one answered question and optional unanswered questions. We simulate this by using the VQA dataset, where each training image $x_i$ is associated with 3 questions $q_{i1}, q_{i2}, q_{i3}$ and their respective answers $a_{i1}, a_{i2}, a_{i3}$. We randomly select a single question per image to be the target question and discard the other answers, leaving $N$ training images $x_i$, each with a question $q_i$, an answer $a_i$ and two additional unanswered questions $q_{i1}'$ and $q_{i2}'$. We train the model on the COCO training image and evaluate on the validation set. Here we use GoogLeNet [39] trained on ILSVRC2012 [36] as the visual representation.

Key experiment: We begin by comparing our iBOWIMG-2x model trained with the extra unanswered questions against the iBOWIMG model of [48] which does not use the available unanswered question. After training our dataset with one answered question per image, iBOWIMG obtains an accuracy of 47.3% on the validation set. In contrast, our model makes effective use of the provided unanswered questions and achieves a significant 3.1% improvement, boosting accuracy to 50.4%. We use bootstrapping to establish statistical significance [14] The 0.999 confidence interval for the baseline model is [46.6%, 48.2%]; thus the improved accuracy of 50.4% when including unanswered questions is statistically significant at the $\alpha = 0.001$ level. Figure 6 demonstrates qualitative results.

Ablation studies: We investigate two components of our model in terms of accuracy improvement: (1) the impact of having access to extra unanswered questions at training time (2) the impact of generating extra training examples per image with data augmentation based on the powerset in Eqn. 2. Table 3 shows the results.

First, as observed above, adding the two additional unanswered questions boosts accuracy by 3.1% from 47.3% to 50.4%. It is further encouraging to note that using just one unanswered question achieves about half the improvement: a 1.6% boost from the 47.3% baseline to 48.9% accuracy with our model. This suggests that adding more unanswered questions (which will become freely available in real-world settings is likely to further improve accuracy.

Second, we investigate the extent of improvement due to data augmentation. Instead of using the data augmentation strategy of Eqn. 2, we simply train iBOWIMG-2x with a single training exemplar $(x_i, q_i, [q_{i1}' q_{i2}'], a_i)$ per image where the two extra questions are concatenated together. This yields an accuracy of 49.2%, which is 1.2% lower than the 50.4% accuracy of the whole model.\footnote{A natural question is whether this improvement arises from seeing the $(x_i, q_i, a_i)$ examples during training since the model is evaluated on test examples of the form $(x, q, \emptyset)$. A model trained with augmentation except without the $(x_i, q_i, \emptyset, a_i)$ examples achieves 50.3% accuracy, indicating this effect is minor.}

This suggests that although most of the improvement comes from simply having access to the extra questions, the fact that the extra questions allow us to generate a diverse augmented training set is in itself a meaningful observation. We explore this further in Section 5.3.

Analysis: Digging deeper, we seek to understand what makes iBOWIMG-2x more effective than iBOWIMG. First, we train a text-only model which learns to answer questions without looking at the image. In this setting, iBOWIMG-2x

<table>
<thead>
<tr>
<th>Unanswered questions</th>
<th>Accuracy w/o aug</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>47.34</td>
<td>47.37</td>
</tr>
<tr>
<td>1 question</td>
<td>48.74</td>
<td>48.94</td>
</tr>
<tr>
<td>2 questions</td>
<td>49.19</td>
<td>50.37</td>
</tr>
</tbody>
</table>

Table 3: Accuracy of iBOWIMG-2x trained with one answered question per image and optional unanswered questions. Models are trained with and without data augmentation of Eqn. 2. The “None w/o aug” setting is equivalent to iBOWIMG [48]. Details in Section 5.1.

\footnote{\textsuperscript{6}Here we evaluate the model in the standard setting where at test-time only the one target question is provided and the model is expected to produce an answer; we do this by inputting a zero-initialized vector as the second textual feature in the model (in place of the unanswered questions).}
achieves an accuracy of 47.3%, which is only a marginal improvement over iBOWIMG's 46.7% accuracy. This suggests that much of the benefit of iBOWIMG-2x is in learning to make better use of the image features. We investigate this further in Section 5.2.

Second, we note that iBOWIMG-2x is more likely to predict answers corresponding to actual words as opposed to a number or yes/no. In particular, iBOWIMG predicts a word answer 72.1% of the time, while iBOWIMG-2x predicts a word answer only 60.2% of the time. Further, iBOWIMG-2x predicts number answers at about half the rate of iBOWIMG: 12.4% compared to 23.7%. This suggests that our model’s richer representation better correlates the image appearance with the semantic textual features, making it more likely to predict a word answer instead of resorting to a simpler numerical or yes/no response.

Table 4 documents the breakdown of accuracy by answer type. Having access to the extra supervisory signal yields a 1.0% improvement on number questions, a bigger 3.0% improvement on the yes/no questions, and a large 3.9% improvement on the challenging word-response questions. Our model is unable to use unanswered questions to learn how to count object instances much better than the baseline; however, it becomes significantly better at identifying the presence or absence of visual concepts and at answering more general visual questions.

Test-time supervision: Finally, an additional advantage of our model is that it can incorporate multiple questions at test time. Concretely, instead of asking a single test question \( q \) on test image \( x \) and passing in the tuple \( (x, q, \emptyset) \) to the model, we consider including other test questions \( q_1' \) and \( q_2' \) and passing in the tuple \( (x, q, [q_1', q_2']) \). This yields an additional 0.5% improvement in accuracy: from 50.4% accuracy (when tested the standard way with only the target question available) to 50.9% accuracy (when all three questions are available simultaneously).

### 5.2. Unanswered questions on novel images

**Dataset:** In Section 5.1 we considered the setting where an answered question is available on every training image. In contrast, here we consider the real-world scenario where some images have only unanswered questions associated with them. To simulate this setting, we randomly select 10% of the training images to be associated with answered questions and we use only unanswered questions on the rest. We evaluate on the full validation set.

**Key experiment:** We use all the available questions to train a visual representation better suited for the VQA task. We use the pretrained AlexNet \([27, 36]\) for the baseline and compare it with the same network finetuned on the COCO training images to recognize 13,759 words from the question vocabulary instead of the 1000 ILSVRC classes. We use these networks as the visual representation when training the iBOWIMG-2x model on the small set of available images with answered questions. The baseline network achieves 43.8% accuracy; the finetuned network effectively utilizes the information captured in the unanswered questions to improve by 1.1% to accuracy of 44.9%.

**Ablation studies:** We evaluate two components of the framework. First, we check whether the full vocabulary is required or if filtering to 80 words (corresponding to the COCO annotated object categories and extracted from the questions as in Section 3.2) or 1024 words (corresponding to the most relevant words according to tf-idf extracted using the code of [48]) would suffice. Figure 7(left) demonstrates continuous improvement with using larger vocabulary sizes. Second, we evaluate whether the full set of unanswered questions is necessary or a smaller subset would suf-
Figure 7: Effectively using training images containing only unanswered questions to improve VQA accuracy by learning the visual representation. The three squares correspond to the same model. Details in Section 5.2.

Benefits of answered vs unanswered questions: We ask one final question: how much does training a better visual representation help compared to collecting more answered questions. In Figure 7(right) we consider progressively increasing the number of available images with answered questions and compare the models with and without finetuning. Interestingly, a model finetuned with only 10% of answered questions achieves an accuracy of 44.9% which is on par with 44.8% accuracy of the model trained on all 100% answered questions without finetuning. This suggests that perhaps much of the information is already captured in the questions themselves even without the answers. However, further study is necessary to verify this claim.

5.3. Data augmentation for VQA

Our findings demonstrate a very simple but effective way of improving VQA accuracy by adding extra unanswered questions. We take this one step further and ask the straightforward question – can we consider the full dataset but use our model as a form of data augmentation, where all questions are used as supervisory signals at training time. Thus, we train iBOWIMG-2x where every image-question-answer triplet is now represented by 8 training exemplars. We use the setup of [48] where the entire COCO training set and 70% of the validation set is used for training. The finetuned GoogLeNet [39] model is used for the visual representation. We evaluate on the test-dev set as standard with only one question at a time provided during testing.

iBOWIMG-2x outperforms the baseline iBOWIMG model by an impressive 7.1%: from 55.7% for iBOWIMG to 62.8% with iBOWIMG-2x having access to the exact same training question-answer pairs but with data augmentation. Table 5 documents the breakdown by answer type. The results are consistent with the findings of Section 5.1; in fact, they are even more pronounced. By making effective use of all the questions jointly through data augmentation, the model improves by 3.1% on the number questions, by 4.2% on the yes/no questions, and by an impressive 10.5% on other questions. This suggests that the data augmentation strategy may be even more beneficial for the open-ended VQA task but we leave that for future work.

These experiments demonstrate that our findings provide important insights not only for the weakly supervised setting but also for the fully supervised VQA scenario. For completeness, our iBOWIMG-2x model achieves 63.17% on test-standard. While this is not state-of-the-art accuracy, the significant 7.1% improvement over the very simple model we started with suggest that our insights may be beneficial for improving the current best models as well.

6. Conclusions

We study a previously unexplored setting of using visual questions themselves as a form of supervision to improve computer vision models. We provide both qualitative and quantitative analysis of how much information is contained within the questions. Our insights already yield significant improvements over baselines on standard benchmarks. More importantly, we believe that visual questions will become freely available as a result of human-AI interactions and can serve as a form of supervision for improving visual models. This work is an early step in this direction.

Acknowledgements

We would like to thank Peiyun Hu, Achal Dave, Arvind Ramachandran, Gunnar Atli Sigurdsson and Siddharth Santurkar for helpful discussions. This research was supported by ONR MURI N000141612007.
References


