

# WeaQA: Weak Supervision via Captions for Visual Question Answering

Pratyay Banerjee   Tejas Gokhale   Yezhou Yang   Chitta Baral

Arizona State University

pbanerj6, tgokhale, yz.yang, chitta@asu.edu

## Abstract

Methodologies for training visual question answering (VQA) models assume the availability of datasets with human-annotated *Image-Question-Answer* (I-Q-A) triplets. This has led to heavy reliance on datasets and a lack of generalization to new types of questions and scenes. Linguistic priors along with biases and errors due to annotator subjectivity have been shown to percolate into VQA models trained on such samples. We study whether models can be trained without any human-annotated Q-A pairs, but only with images and their associated textual descriptions or captions. We present a method to train models with synthetic Q-A pairs generated procedurally from captions. Additionally, we demonstrate the efficacy of spatial-pyramid image patches as a simple but effective alternative to dense and costly object bounding box annotations used in existing VQA models. Our experiments on three VQA benchmarks demonstrate the efficacy of this weakly-supervised approach, especially on the VQA-CP challenge, which tests performance under changing linguistic priors.

## 1 Introduction

Since Visual Question Answering (VQA) was first proposed as a Turing test (Malinowski and Fritz, 2014), several human-annotated datasets (Mogadala et al., 2019) have been used to train and evaluate VQA models. Unfortunately, heavy reliance on these datasets for training has the unwanted side-effects of bias towards answer styles, question-types (Chao et al., 2018), and spurious correlations with language priors (Agrawal et al., 2018). Similar findings have been reported for natural language tasks (Gururangan et al., 2018; Niven and Kao, 2019; Kaushik et al., 2020). Evaluating VQA models on test-sets that are very similar to training sets is deceptive and inadequate and not an accurate measure of robustness.

To address this, one line of work has focused on balancing, de-biasing, and diversifying samples (Goyal et al., 2017; Zhang et al., 2016). However, crowd-sourcing “unbiased” labels is difficult and costly; it requires a well-designed annotation interface and a large-scale annotation effort with dedicated and able annotators (Sakaguchi et al., 2020). The alternative (that this paper aligns itself with) is to avoid the use of explicit human annotations and instead to train models in an unsupervised manner by synthesizing training data. These techniques, coined *unsupervised*<sup>1</sup>, come with many advantages – human bias and subjectivity are reduced; the techniques are largely domain-agnostic and can be transferred from one language to another (low resource languages) or from one visual domain to another. For instance, template-based Q-A generation developed for synthetic blocks-world images in CLEVR (Johnson et al., 2017) can also be used to generate Q-A pairs for natural complex scenes in GQA (Hudson and Manning, 2019) or the referring-expressions task (Liu et al., 2019).

In this work, we train VQA models without using human-annotated Q-A pairs. Instead, we rely on weak supervision from image-captioning datasets, which provide multi-perspective, concise, and less subjective descriptions of visible objects in an image. We procedurally generate Q-A pairs from these captions and train models using this synthetic data, and *only evaluate* them on established human-annotated VQA benchmarks.

**Why Captions?** Image captioning, like VQA, has been a central area of vision-and-language research. Datasets such as MS-COCO (Lin et al., 2014; Chen et al., 2015) contain captions that describe objects and actions in images of everyday scenes. During the construction of MS-COCO, human captioners were instructed to refrain from describing past and future events or “what a person might say”. On the other hand, annotators of

VQA (Antol et al., 2015) were instructed to ask questions that “a smart robot cannot answer, but a human can” and “interesting” questions that may require “commonsense”. Different sets of annotators provided answers to these questions and were allowed to speculate or even guess an answer that *most people would agree on*. It has also been shown that multiple answers may exist for questions in common VQA datasets (Bhattacharya et al., 2019).

In Figure 2, the first VQA-v2 question asks how many doors the car has. Although commonsense (and linguistic priors) would suggest that “Most cars have *four* doors”, only two doors can be seen in the image. What should the model predict, *two* or *four*? The second question is subjective and has multiple contradicting answers from different annotators (where one should draw the line between opaque, transparent, or reflective is not very clear). Similarly, the first GQA question is ambiguous and could refer to either the skier or the photographer.

Thus the very nature of the data-collection procedure and instructions for VQA brings in human subjectivity and linguistic bias as compared to caption annotations, which are designed to be simple, precise, and non-speculative. Motivated by this, we study the benefits of using captions to synthesize Q-A pairs, using three types of methods:

1. template-based methods similar to (Ren et al., 2015a; Gokhale et al., 2020b),
2. paraphrasing and back-translation (Sennrich et al., 2016) which provide linguistic variation,
3. synthesis of questions about image semantics using the QA-SRL (He et al., 2015) approach.

Since our Q-A pairs are created synthetically, there does exist a domain shift as well as label (answer) shift from evaluation datasets such as VQA-v2 and GQA as shown in Figure 2, thus posing challenges to this weakly-supervised method.

We evaluate two models, UpDown (Anderson et al., 2018) and a transformer-encoder (Vaswani et al., 2017) based model pre-trained on synthetic Q-A pairs and image-caption matching task. To remove the dependence on object bounding-boxes and labels needed to extract object features, we propose spatial pyramids of image patches as a simple and effective alternative.

To the best of our knowledge, this is the first work on the unsupervised<sup>1</sup> visual question answering, with the following contributions:

<sup>1</sup>adhering to the usage of this term in Lewis et al. (2019a).

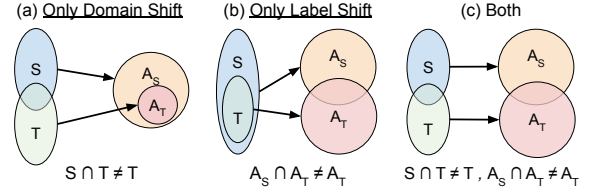


Figure 1: Aspects of generalization in VQA.

- We introduce a framework for synthesizing (*Question*, *Answer*) pairs from captions.
- Since synthetic samples (unlike popular benchmarks) include multi-word answer phrases, we propose a sub-phrase weighted-answer loss to mitigate bias towards such multi-word answers.
- We propose pre-training tasks that use spatial pyramids of image-patches instead of object bounding-boxes, further removing the dependence on human annotations.
- Extensive experiments and analyses under zero-shot transfer and fully-supervised settings on VQA-v2, VQA-CP, and GQA show our model’s efficacy and establish a strong baseline for future work on unsupervised visual question answering.

## 2 Related Work

**Robustness in VQA** can be defined as shown in Figure 1 under two situations: domain shift and label shift. Under domain shift, generalization to a new input domain (such as different styles of questions or novel scenes) is desired, characterized by  $S \cap T \neq T$  where  $S$  and  $T$  denote the train and test input domains. Under label shift, generalization to novel answers is desired (predicting answers not seen during training), characterized by  $A_S \cap A_T \neq A_T$ , where  $A_S$  and  $A_T$  are the set of answers seen during training and test-time.

Performance under **domain shift** has been evaluated for new domains of test questions with unseen words and objects (Teney and Hengel, 2016; Ramakrishnan et al., 2017), novel compositions (Johnson et al., 2017; Agrawal et al., 2017), logical connectives (Gokhale et al., 2020b), as well as questions that are implied (Ribeiro et al., 2019), entailed (Ray et al., 2019) or sub-questions (Selvaraju et al., 2020); or for datasets with varying linguistic styles (Chao et al., 2018; Xu et al., 2020; Shrestha et al., 2019) and different reasoning capabilities (Kafle and Kanan, 2017).

**Label shift** or Prior Probability Shift (Storkey,



Captions		Question	Answer(Confidence)
<ul style="list-style-type: none"> <li>- A car that seems to be parked illegally behind a legally parked car</li> <li>- A couple of cars parked in a busy street sidewalk</li> <li>- Cars try to maneuver into parking spaces along a densely packed street.</li> <li>- two cars parked on the sidewalk on the street</li> </ul>		<b>VQA-v2</b> 1. How many doors does the gray car have ? 2. Why does the windshield look opaque ?	4 (1.0) Clear (0.6), No (0.3), Reflection (0.9)
<ul style="list-style-type: none"> <li>- A man in skis is coming up the hill</li> <li>- A skier is passing a competition race marker</li> <li>- A man takes a picture of a skier</li> <li>- A cross-country skier is competing at night in snow</li> </ul>		<b>GQA</b> 1. Is the man on the left or on the right ? 2. Who is wearing the jersey ? 3. What is someone passing ? 4. When is someone competing ? 5. Who is coming ? 6. Is that a man in skateboard coming up the hill ? 7. Where is someone coming ?	Right (1.0) Man (1.0) A competition race marker (1.0) At night (1.0) A man in skis (1.0) No Up the hill (1.0)

Figure 2: Examples of images and human-annotated Q-A pairs from VQA and GQA and our synthetic Q-A pairs.

2009) has been implicitly explored in VQA-CP (Agrawal et al., 2018), where the conditional probabilities of answers given the question type deviate at test-time. Teney et al. (2020c) have identified several pitfalls associated with the models and evaluation criteria for VQA-CP.

**Unsupervised Extractive QA** in which aligned (*context, question, answer*) triplets are not available, has been studied (Lewis et al., 2019b; Banerjee and Baral, 2020; Rennie et al., 2020; Fabbri et al., 2020; Li et al., 2020; Banerjee et al., 2021) by training models on procedurally generated Q-A pairs. Captions have been used to generate Q-A pairs for logical understanding (Gokhale et al., 2020b) and commonsense video understanding (Fang et al., 2020a). Li et al. (2018); Krishna et al. (2019) have explored Visual Question Generation from an input image and answer.

**Weak supervision** is an active area of research; for instance in action/object localization (Song et al., 2014; Zhou et al., 2016) and semantic segmentation (Khoreva et al., 2017; Zhang et al., 2017) without pixel-level annotations, but only class labels. There is also interest growing in leveraging natural language captions or textual queries as weak supervision for visual grounding tasks (Hendricks et al., 2017; Mithun et al., 2019; Fang et al., 2020b).

**Visual Feature Extractors** such as VGG (Simonyan and Zisserman, 2015) and ResNet (He et al., 2016) have been widely used for many computer vision tasks. Object-based features such as RCNN (Girshick et al., 2014) and Faster-RCNN (Ren et al., 2015b) have become the standard for V&L tasks (Anderson et al., 2018).

### 3 Framework for Synthesizing Q-A Pairs

**Problem Statement:** Consider a dataset containing images and associated captions as shown in Figure 2. Our work deals with learning VQA using these image-caption data, without any labeled Q-A pairs, and answer questions about unseen images.

#### 3.1 Question Generation

Several studies (Du et al., 2017; Lewis et al., 2019a) have been dedicated to the complex domain of question generation. We approach it conservatively, using template-based methods and semantic role labeling, with paraphrasing and back-translation for improving the linguistic diversity of template-based questions. We begin by extracting object words from the caption by using simple heuristics such as extracting noun-phrases and using numerical quantifiers in the caption as soft approximations of objects’ cardinality. If object-words are available explicitly, we used them as is. Questions are categorized based on answer types; *Yes-No*, *Number*, *Color*, *Location*, *Object*, and *Phrases*.

**Template-based:** To create *Yes-No* questions, modal verbs are removed from the caption, and a randomly chosen question prefix such as “*is there*”, “*is this*” is attached. For instance, the caption “A man is wearing a hat and sitting” is converted to “*Is there a man wearing a hat and sitting*”, with the answer “Yes”. To create the corresponding question with the answer “No”, we use either negation or replace the object-word with an adversarial word or antonym, thus obtaining “*Is there a dog wearing a hat and sitting*” for which the answer is “No”. An adversarial word refers to an object absent in the image but similar to objects in the image. To compute similarity, we use Glove (2014) word-vectors.

For *Object*, *Number*, *Location*, and *Color* questions, we follow a procedure similar to Ren et al.



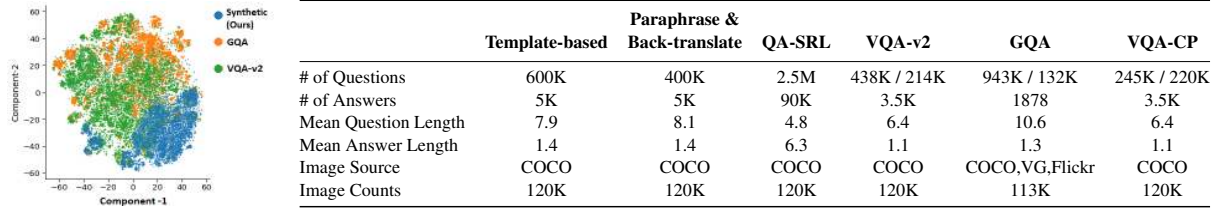


Figure 3: Discrepancy between VQA-v2, GQA, and synthetic samples. Left: t-SNE plot of question embeddings. Right: Dataset statistics for our generated Q-A pairs with Train/Val. splits for benchmark datasets.

(2015a). To create “*what*” questions for the *Object* type, we extract objects and noun phrases from captions as potential answers and replace them with *what*. The question is rephrased by splitting long sentences into shorter ones and converting indefinite determiners to definite. A similar procedure is used for *Number* questions; numeric quantifiers of noun phrases are extracted and replaced by “how many” and “what is the count” to form the question. *Color* questions are generated by locating the color adjective and the corresponding noun phrase and replacing them in a templated question: “What is the color of the object?”. *Location* questions are similar to *Object* questions, but we extract phrases with “in”, “within” to extract locations, with places, scenes, and containers as answers.

**Semantic Role Labeling:** QA-SRL (He et al., 2015) was proposed as a paradigm to use natural language to annotate data by using Q-A pairs to specify textual arguments and their roles. Consider the caption “A girl in a red shirt holding an apple sitting in an empty open field”. Using QA-SRL with B-I-O span detection and sequence-to-sequence models (FitzGerald et al., 2018), for the “*when*”, “*what*”, “*where*”, and “*who*” questions, we obtain Q-A pairs belonging to the *Phrases* category such as:

(what is someone holding?, an apple)  
 (who is sitting?, girl in a red shirt holding an apple)  
 (where is someone sitting?, an empty open field)

These examples illustrate that QA-SRL questions are short and use generic descriptors such as *something* and *someone* instead of elaborate references, while the expected answer phrases are longer and descriptive. Thus to answer these, better semantic image understanding is required.

**Paraphrasing and Back-Translation (P&B):** We apply two natural language data augmentation techniques, paraphrasing, and back-translation to increase the linguistic variation in the questions. To paraphrase questions, we train a T5 (Raffel et al.,

2019) text generation model on the Quora Question Pairs Corpus (). For back-translation, we train another T5 text generation model on the Opus corpus (2012), translate the question to an intermediate language (Français, Deutsche, or Español), and translate the question back to English. For example:

Is the girl who is to the left of the sailboats wearing a backpack?  
 ↓ Español  
 La chica que está a la izquierda de los veleros lleva mochila?  
 ↓ English  
 Does the girl to the left of the sailboats carry a backpack?

### 3.2 Domain Shift w.r.t. VQA-v2 and GQA

Compared to current VQA benchmarks (which typically contain one-word answers), answers to QA-SRL questions are more descriptive and contain adjectives, adverbs, determiners, and quantifiers, as seen in Figure 2. On the other hand, synthetic questions have less descriptive subjects due to the use of pronouns. Our synthetic data contains 90k unique answer phrases, compared to 3.2k in VQA and 3k in GQA. Around 200 answers from VQA are not present in our answer phrases, such as time (11:00) and proper nouns (LA Clippers), both of which are not present in caption descriptions.

Moreover, our training data contains Q-A pair such as (“Where is the man standing?”, “to the left of the table”), generated by QA-SRL with long phrases as answers. However, the test set contains questions such as (“Which side of the car is the tree?”, “left”), which expects only “left” as the answer. So although the word “left” is seen as a sub-phrase of our training answers, it is not explicitly seen as an only correct answer.

Some of our synthetic template-based questions about counting and object presence are similar in style to those in VQA and GQA. However, QA-SRL questions require a semantic understanding of the actions depicted in the image, which are rare in VQA and GQA. We quantify this by plotting the t-SNE components of document vector embed-

dings of the questions from VQA, GQA, and our synthetic data, in Figure 3, and observe that our synthetic questions are a distinct cluster, while VQA and GQA overlap with each other. As such, a linguistic domain shift exists between these synthetic source questions and human-annotated target questions. In this paper, we address the challenge of learning VQA on a synthetically generated dataset and evaluating models on conventional benchmarks which have questions and answers that deviate linguistically from synthetic training samples.

## 4 Method

Recently, multiple deep transformer-based architectures have been proposed (Tan and Bansal, 2019; Lu et al., 2019; Chen et al., 2019), that are pre-trained on a combination of multiple VQA and image captioning datasets such as Conceptual Captions (Sharma et al., 2018), SBU Captions (Ordonez et al., 2011), Visual Genome (Krishna et al., 2017), and MSCOCO (Lin et al., 2014). These models are resource intensive as they are trained on a huge collection of data with 3 million images. We train our models only on MS-COCO captions and images ( $\sim 204k$ ), without access to any human-authored Q-A pairs or object bounding boxes.

### 4.1 Spatial Pyramid Patches

“Bottom-Up” object features (Anderson et al., 2018) extracted from Faster R-CNN (Ren et al., 2015b) have become the de-facto features used in state-of-the-art VQA models. These VQA models thus only use features of detected objects as input, and ignore the rest of the image. Although object features are discriminative, dense annotations are required for training and additional large deep networks for extraction. Object detection can be imperfect for small and rare objects (Wang et al., 2019); for instance if an object detection model detects only four out of six bananas in an image, features of the other two bananas will not be used by VQA models. This creates a performance bottle-neck for questions about counting or rare objects.

We take a step back and postulate that the use of features of the entire image in context could reduce this bottleneck. Image features extracted from a ResNet (He et al., 2016) trained for the ImageNet (Russakovsky et al., 2015) classification task, which is widely used for computer vision tasks, have been previously used for VQA models (Goyal et al., 2017). Unfortunately, since Im-

geNet contains iconic (single-object) images, using these features for non-iconic VQA images is restrictive since many questions refer to multiple objects and backgrounds in the image. Inspired by Spatial Pyramid Matching (Lazebnik et al., 2006) for image classification, we propose *spatial pyramid patch features* to represent the input VQA image into a sequence of features at different scales.

We divide each image  $I$  into a set of image patches  $\{I_{k_1}, \dots, I_{k_n}\}$ , each  $I_{k_i}$  being a  $k_i \times k_i$  grid of patches, and extract ResNet features for each patch. Larger patches encode global features and relations, while smaller patches encode local and low-level features.

**Encoder:** Our Encoder model is similar to the UNITER single-stream transformer, where the sequence of word tokens  $w = \{w_1, \dots, w_T\}$  and the sequence of image patch features  $v = \{v_1, \dots, v_K\}$  are taken as input. We tokenize the text using a WordPieces (Wu et al., 2016) tokenizer similar to BERT (Devlin et al., 2019), and embed the text tokens through a text-embedder (Sanh et al., 2019). The visual features are projected to a shared embedding space using a fully-connected layer. A projected visual position encoding, indicating the patch region (top-right, bottom-left) is added to the visual features. We concatenate both sequences of features and feed them to  $L$  cross-modality attention layers. Parameters between the cross-modality attention layers are shared to reduce parameter count and increase training stability (Lan et al., 2020), and a residual connection and layer normalization is added after cross-modal attention layer similar to Vaswani et al. (2017).

### 4.2 Pre-training Tasks and Loss Functions

We train the Encoder model using three pre-training tasks: Masked Language Modeling, Masked Question Answering, and Image-Text Matching.

**Masked Language Modeling (MLM):** We randomly mask 15% of the word tokens from the caption and ask the model to predict them. For the caption “There is a man wearing a hat”, the model gets the input “There is [MASK] wearing a hat”. Without the image, there can be multiple plausible choices for the [MASK] token, such as “woman”, “man”, “girl”, but given the image the model should predict “man”. This task has been shown to effectively learn cross-modal features (2019).

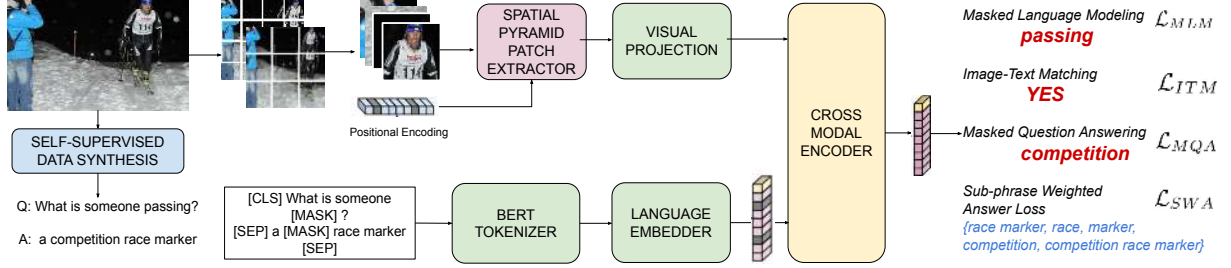


Figure 4: Our model architecture makes the use of spatial pyramids of image patches as inputs to the Encoder, which is trained for three pre-training tasks as shown.

**Masked Question Answering (MQA):** In this task, the answer tokens are masked, and the model is trained to predict the answer tokens. For example in Figure 2, for the input “When is someone competing? [MASK] [MASK]”, the model should predict, “at night”. To answer such questions, the model needs to interpret the image.

**Image-Text Matching (ITM):** We use the five captions provided by MS-COCO as positive samples for each image. To obtain negative samples, we randomly sample captions from other images that contain a different set of objects. We train the model on a binary classification task (matching / not matching) for each image-caption pair.

For VQA and ITM, we use the final layer representation  $z^{[CLS]}$  of [CLS] token, followed by a feed-forward and softmax layer. For MLM and MQA we feed corresponding token representations to a different feed-forward layer. We train the model using cross-entropy loss for all three tasks.

**Sub-phrase Weighted Answer Loss:** As observed before, the questions generated in QA-SRL have long answer phrases. For instance “What is parked?” has the answer “two black cars”. We extract all possible sub-phrases that can be alternate answers, but assign them a lower weight than the complete phrase, computed as  $W_{sub} = \text{WordCount}(sub) / \text{WordCount}(ans)$ . Thus “two black cars” has a weight 1.0, while the extracted sub-phrases and weights are: (two, 0.33), (2, 0.33), (black, 0.33), (cars, 0.33), (two cars, 0.66), (2 cars, 0.66), (black cars, 0.66), (car, 0.33). This enforces a distribution over the probable answer space instead of a strict “single true answer” training. We train the model with this additional binary cross-entropy loss, where the model predicts a weighted distribution  $y_{wa}$  over the answer vocabulary. The vocabulary is defined from the synthetic

QA answer-space.

$$\mathcal{L}_{SWA} = \mathcal{L}_{BCE}(\sigma(z^{[CLS]}), y_{wa}). \quad (1)$$

The total loss, with scalar coefficients  $\alpha, \beta \in (0, 1]$  is given by:

$$\mathcal{L} = \mathcal{L}_{MLM} + \mathcal{L}_{MQA} + \alpha \cdot \mathcal{L}_{ITM} + \beta \cdot \mathcal{L}_{SWA}. \quad (2)$$

## 5 Experimental Setup

**Datasets:** We evaluate our methods on the three popular visual question answering benchmarks: VQA-v2, VQA-CP-v2, and GQA. Answering questions in VQA-v2 and VQA-CP v2 requires image and question understanding, whereas GQA further requires spatial understanding such as compositionality and relations between objects. We evaluate our methods under *zero-shot* transfer (trained only on procedurally generated samples), and *fully-supervised* (where we finetune our model using the associated train annotations) settings. We use exact-match accuracies for GQA, and use VQA-metric (Agrawal et al., 2017) for VQA.

**Training:** Our Encoder has 8 cross-modal layers with a hidden dimension of 768. The weights are initialized using the standard definition as provided in the Huggingface repository (Wolf et al., 2019). Our models are pre-trained for 40 epochs with a learning rate of  $1e-5$ , batch size of 256, using Adam optimizer. For finetuning, we use a learning rate of  $1e-5$  or  $5e-5$  and batch size of 32 for 10 epochs. We use a ResNet-50 pretrained on ImageNet to extract features from image patches with 50% overlap, and Faster R-CNN pretrained on Visual Genome to extract object features. We evaluate both frozen and finetuned ResNet, and observe finetuning the feature extractor to perform better. All our models are trained using 4 Nvidia V100 16 GB GPUs. All results in the fully supervised setting are reported for from-scratch trained final classification layers.

Model	All	Yes-No	Num	Others
SAN (2016)	25.0	38.4	11.1	21.7
GVQA (2018)	31.3	58.0	13.7	22.1
UpDown (2018)	39.1	62.4	15.1	34.5
AReg(2017)	42.0	65.5	15.9	36.6
AdvReg (2019)	42.3	59.7	14.8	40.8
RUBi (2019)	47.1	68.7	20.3	43.2
Teney and van den Hengel (2019)	46.0	58.2	29.5	44.3
Unshuffling (2020b)	42.4	47.7	14.4	47.3
UpDn+CE+GS (2020a)	46.8	64.5	15.4	45.9
LXMERT (2019)	46.2	42.8	18.9	55.5
SCR (2019)	48.4	70.4	10.4	47.3
LMH (2019)	52.4	69.8	44.5	45.5
CSS (2020)*	58.9	84.4	49.4	48.2
MUTANT (2020a)*	<b>69.5</b>	<b>93.2</b>	<b>67.2</b>	<b>57.8</b>
ZSL+Objects+UpDown	40.8	67.4	28.6	30.2
ZSL+Patches+UpDown	41.2	68.5	29.8	30.0
ZSL+Patches+Encoder	<u>47.3</u>	<u>73.4</u>	<u>39.8</u>	<u>35.6</u>

Table 1: Unsupervised accuracy on VQA-CP-v2 test set. All baselines are *supervised* methods trained on the train split. \* use further additional supervised training samples. Cyan: our model is better overall. Red: our model is better on specific categories.<sup>2</sup>

Model	All	Yes-No	Num	Others
GVQA (2018)	48.2	72.0	31.1	34.7
UpDown (2018)	65.3	81.8	44.2	56.1
RUBi (2019)	63.1	*	*	*
MCAN (2019)	70.4	85.8	53.7	60.7
ViBERT (2019)	70.5	*	*	*
LXMERT (2019)	72.5	<b>88.2</b>	<b>54.2</b>	<b>63.1</b>
UNITER (2019)	<b>72.7</b>	*	*	*
ZSL + Objects + UpDown	41.4	68.1	27.6	29.4
ZSL + Patches + UpDown	40.6	67.8	28.4	29.2
ZSL + Patches + Encoder	<u>46.8</u>	<u>72.1</u>	<u>34.4</u>	<u>34.1</u>
FSL + Objects + UpDown	66.8**	82.4**	45.1**	56.4**
FSL + Patches + UpDown	63.4	80.2	45.2	52.1
FSL + Patches + Encoder	65.3	80.5	48.94	56.2

Table 2: VQA-v2 Test-standard accuracies<sup>2</sup>. FSL models are pretrained on synthetic samples, and further finetuned on VQA-v2 train split. \* - Scores are not available, \*\* - Validation split scores.

**Baselines:** To measure the improvements due to our proposed image patch features and SWA loss, we compare our methods to the UpDown model Anderson et al., which uses object bounding-box features. For the Zero-shot transfer setting, we compare our Encoder with UpDown when trained with spatial features as well as object features. Pre-trained transformers such as UNITER use large V&L corpora, dense human annotations for objects and Q-A pairs and supervised loss functions over these. Comparisons with such models are therefore not fair in a ZSL setting; instead, we perform these comparisons in a fully-supervised (FSL) setting.

<sup>2</sup>ZSL refers to zero-shot transfer setting and FSL refers to our models further finetuned on the respective train split. Underline⇒unsupervised best, **bold**⇒overall best. Baselines are trained on train-split, our models on synthetic data.

Model	All	Binary	Open
CNN + LSTM (2018)	46.6	61.9	22.7
UpDown (2018)	49.7	66.6	34.8
MAC (2018)	54.1	71.2	38.9
BAN (2018)	57.1	76.0	40.4
LXMERT (2019)	<b>60.3</b>	<b>77.8</b>	<b>45.0</b>
ZSL + Objects + UpDown	30.7	50.8	17.6
ZSL + Patches + UpDown	31.1	52.3	16.8
ZSL + Patches + Encoder	<u>33.7</u>	<u>55.5</u>	<u>21.2</u>
FSL + Objects + UpDown	50.4	67.5	35.1
FSL + Patches + UpDown	46.4	64.3	31.4
FSL + Patches + Encoder	55.2	73.6	38.8

Table 3: GQA Validation split accuracies.<sup>2</sup>

## 6 Results<sup>2</sup>

**Unsupervised Question Answering:** Tables 1, 2 and 3 summarize our results on the three benchmark datasets. We can observe that our method outperforms specially designed supervised methods for bias removal in VQA-CP; our model with UpDown is 1.1% better than the supervised UpDown. Under the ZSL setting for VQA-CP, our Encoder model is 6.1% better than UpDown with patches, and 6.5% better than UpDown with Object features, for VQA-v2: 6.2%, 5.4% respectively, and for GQA: 2.2%, 3.0% respectively.

For VQA-CP, our procedurally generated Q-A pairs and patch-features when used with either UpDown or Encoder are better than the baseline supervised UpDown model, showing the improvements are model-agnostic. This also shows the merits of using our Q-A generation methods when train and test-sets deviate linguistically.

Most GQA questions require understanding spatial relationships between objects. Such questions are infrequent in our synthetic training data since captions do not contain detailed spatial relationships among objects. Thus, the ZSL performance is not as competitive for GQA when compared to our performance on VQA and VQA-CP. Improving spatial and compositional question-answering with weak supervision is an interesting future pursuit.

**Fully Supervised Question Answering:** In the FSL setting, our methods’ performance is not far from SOTA methods, even though our method uses significantly fewer annotations (no access to object bounding boxes). In GQA, the Encoder model performs on par with MAC (2018) and BAN (2018), which unlike us, use object relationship annotations. This suggests that cross-modal transformer layers can learn spatial relations from spatial pyramidal



	Question Generation	VQA-v2	VQA-CP	GQA
UpDn	Template	26.2	25.7	11.6
	Template + Para&Back	28.5	27.1	14.8
	QA-SRL	31.1	33.8	18.9
	All	41.4	40.2	31.1
Encoder	Template	32.5	31.3	18.5
	Template + Para&Back	34.8	33.6	23.6
	QA-SRL	40.3	39.8	21.4
	All	<b>47.1</b>	<b>46.8</b>	<b>33.7</b>

Table 4: Effect of different pre-training data sources on ZSL Validation split accuracies.

	Patch Resolutions	VQA-v2	VQA-CP	GQA
UpDn	{1}	18.8	19.7	11.3
	{1, 3}	36.7	35.9	24.5
	{1, 3, 5}	40.1	39.7	29.5
	{1, 3, 5, 7}	<b>41.4</b>	<b>40.2</b>	<b>31.1</b>
	{1, 3, 5, 7, 9}	39.8	38.4	29.3
Encoder	{1}	26.4	27.7	15.3
	{1, 3}	42.6	43.1	28.8
	{1, 3, 5}	44.3	45.2	30.9
	{1, 3, 5, 7}	<b>47.1</b>	<b>46.8</b>	<b>33.7</b>
	{1, 3, 5, 7, 9}	46.2	45.4	31.2

Table 5: Effect of the number of spatial patches on ZSL performance {3,5} implies division of the image into a 3x3 and 5x5 grid of patches.

features.

**Impact of each question-generation technique:** In Table 4 we can observe the effect of different question generation techniques. All models use spatial image patch features. QA-SRL based questions and the SWA-Loss contribute the most towards gains in performance, and the paraphrased questions provide larger linguistic variation.

**Effect of Spatial Pyramids:** We study the effect of progressively increasing the number of spatial image patches (i.e., decreasing the patch size). Table 5 shows that an optimum exists at grid-size of  $7 \times 7$  after which the addition of smaller patches is detrimental. Similarly, only using patches of large size does not allow models to focus on specific image regions. Thus a trade-off exists between global context and region-specific features. Changing the feature extractor from ResNet-50 to ResNet-101 only results in a minor improvement of 0.01% to 0.30%. Removing visual position embeddings has a significant effect on performance, with a drop of 4.60% to 8.00% in both ZSL and FSL settings.

**Impact of Pre-training Tasks:** Table 6 shows the effect of different pretraining tasks on the downstream zero-shot transfer VQA task. We need the

Pre-Training Task	VQA-v2	VQA-CP	GQA
SWA	39.1	38.3	25.4
MLM+SWA	42.4	41.5	27.8
MQA+SWA	42.0	41.2	26.6
MLM+MQA+SWA	45.6	44.9	29.7
MLM+ITM+SWA	44.7	43.6	28.9
<b>All</b>	<b>46.2</b>	<b>45.4</b>	<b>31.2</b>

Table 6: Effect of different pre-training tasks on the ZSL performance for the Encoder model.

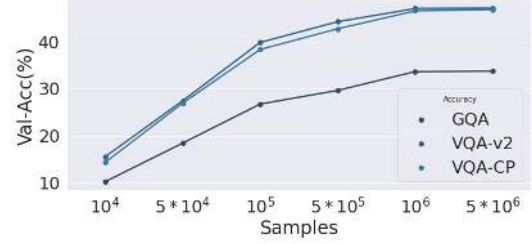


Figure 5: Learning Curve showing validation accuracy vs. number of synthetically generated training samples.

SWA task, as it is used to perform the zero-shot QA task. The combination of MLM, MQA, and ITM, all of which need image understanding, shows improved performance on the downstream task, indicating better cross-modal representations.

**Effect of size of synthetic training set:** Figure 1 shows our Encoder model’s learning curve for the zero-shot transfer setting trained on our synthetic Q-A pairs. The performance stagnates after a critical threshold of  $10^6$  samples is reached. Our experiments also suggest that randomly sampling a set of questions for each image per epoch leads to a 4% gain compared to training on the entire set.

**Error Analysis:** Our ZSL method is pretrained on longer phrases and hence tends to generate more detailed answers, such as “red car” instead of “car”. Although the SWA loss is designed to encourage a distribution over the shorter phrases, the bias is not entirely removed. On automated evaluation, we observe that for 42% of questions, the target answer is a sub-phrase of our predicted answer. Manual evaluation of 100 such samples shows that 87% of such detailed predicted answers are plausible. This shows the relevance of learning from captions and quantifies the bias towards short “true” answers in human-annotated benchmarks, calling for better evaluation metrics that do not penalize VQA systems for producing descriptive or alternative accurate answers.

In the FSL setting, we either finetune our pre-



trained QA classifier with the SWA Loss or train a separate feedforward layer from scratch for the task. The pre-trained QA classifier predicts longer phrases as answers, leading to a drop in accuracy. The feedforward layer performs better (+6%), indicating our Encoder captures relevant features necessary to generalize to the benchmark answer-space. Note that we do not use object annotations during training, unlike existing methods.

Our error analysis and Figure 3 show the shift in question-space and answer-space between synthetic and human-authored Q-A pairs. These (along with inadequate evaluation metrics) act as the primary sources explaining the performance-gap between weakly-supervised methods and the fully-supervised setting. It remains to be seen whether more sophisticated question generation can be developed to reduce the performance gap further and mitigate the heavy reliance on human annotations.

## 7 Discussion and Conclusion

Prior work (Chen et al., 2019; Jiang et al., 2020) has demonstrated that the use of object bounding-boxes and region features leads to significant improvements on downstream tasks such as captioning and VQA. However, little effort has been dedicated to developing alternative methods that can approach similar performance without relying on dense annotations. We argue that weakly supervised learning coupled with data synthesis strategies could be the pathway for the V&L community towards a “post-dataset era”.<sup>2</sup> In this work, we take a step towards that goal. We address the problem of weakly-supervised VQA with a framework for the procedural synthesis of Q-A pairs from captions for training VQA models, where benchmark datasets can be used only for evaluation. We use spatial pyramids of patch features to increase the annotation efficiency of our methods. Our experiments and analyses show the potential of patch-features and procedural data synthesis and reveal problems with existing evaluation metrics.

## Ethical Considerations

Captions and Question-Answer pairs are both annotated by humans in existing image captioning and visual question answering datasets. However, captions arguably contain a lesser degree of subjectivity, ambiguity, and linguistic biases than VQA annotations, due to the design of annotation prompts

that limit the introduction of these biases. Our work points to the potential of procedurally generated annotations in providing robustness improvements under changing linguistic priors in VQA test sets (Table 1). Hendricks et al. find that gender bias exists in image-captioning datasets and is *amplified* by models; further research in self-supervised data synthesis could potentially help alleviate such social biases.

## Acknowledgements

The authors acknowledge support from the DARPA SAIL-ON program W911NF2020006, ONR award N00014-20-1-2332, and NSF grant 1816039, and the anonymous reviewers for their insightful discussion.

## References

- Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. [Don’t just assume; look and answer: Overcoming priors for visual question answering](#). In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 4971–4980. IEEE Computer Society.
- Aishwarya Agrawal, Aniruddha Kembhavi, Dhruv Batra, and Devi Parikh. 2017. C-vqa: A compositional split of the visual question answering (vqa) v1.0 dataset. *arXiv preprint arXiv:1704.08243*.
- Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. [Bottom-up and top-down attention for image captioning and visual question answering](#). In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6077–6086. IEEE Computer Society.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. [VQA: visual question answering](#). In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433. IEEE Computer Society.
- Pratyay Banerjee and Chitta Baral. 2020. [Self-supervised knowledge triplet learning for zero-shot question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 151–162, Online. Association for Computational Linguistics.
- Pratyay Banerjee, Tejas Gokhale, and Chitta Baral. 2021. [Self-supervised test-time learning for reading](#).

<sup>2</sup>A. Efros, *Imagining a post-dataset era*, ICML’20 Talk.

- comprehension. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1200–1211, Online. Association for Computational Linguistics.
- Nilavra Bhattacharya, Qing Li, and Danna Gurari. 2019. [Why does a visual question have different answers?](#) In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4270–4279. IEEE.
- Rémi Cadène, Corentin Dancette, Hedi Ben-younes, Matthieu Cord, and Devi Parikh. 2019. [Rubi: Reducing unimodal biases for visual question answering.](#) In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 839–850.
- Wei-Lun Chao, Hexiang Hu, and Fei Sha. 2018. [Cross-dataset adaptation for visual question answering.](#) In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5716–5725. IEEE Computer Society.
- Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. 2020. [Counterfactual samples synthesizing for robust visual question answering.](#) In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10797–10806. IEEE.
- Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.
- Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740*.
- Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. [Don’t take the easy way out: Ensemble based methods for avoiding known dataset biases.](#) In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4069–4082, Hong Kong, China. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding.](#) In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Xinya Du, Junru Shao, and Claire Cardie. 2017. [Learning to ask: Neural question generation for reading comprehension.](#) In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1342–1352, Vancouver, Canada. Association for Computational Linguistics.
- Alexander Fabbri, Patrick Ng, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. [Template-based question generation from retrieved sentences for improved unsupervised question answering.](#) In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4508–4513, Online. Association for Computational Linguistics.
- Zhiyuan Fang, Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020a. [Video2Commonsense: Generating commonsense descriptions to enrich video captioning.](#) In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 840–860, Online. Association for Computational Linguistics.
- Zhiyuan Fang, Shu Kong, Zhe Wang, Charles Fowlkes, and Yezhou Yang. 2020b. Weak supervision and referring attention for temporal-textual association learning. *arXiv preprint arXiv:2006.11747*.
- Nicholas FitzGerald, Julian Michael, Luheng He, and Luke Zettlemoyer. 2018. [Large-scale QA-SRL parsing.](#) In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2051–2060, Melbourne, Australia. Association for Computational Linguistics.
- Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. [Rich feature hierarchies for accurate object detection and semantic segmentation.](#) In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 580–587. IEEE Computer Society.
- Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020a. [MUTANT: A training paradigm for out-of-distribution generalization in visual question answering.](#) In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 878–892, Online. Association for Computational Linguistics.
- Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020b. [Vqa-lol: Visual question answering under the lens of logic.](#) In *European Conference on Computer Vision (ECCV)*.

- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. [Making the V in VQA matter: Elevating the role of image understanding in visual question answering](#). In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society.
- Gabriel Grand and Yonatan Belinkov. 2019. [Adversarial regularization for visual question answering: Strengths, shortcomings, and side effects](#). In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, pages 1–13, Minneapolis, Minnesota. Association for Computational Linguistics.
- Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. [Annotation artifacts in natural language inference data](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. [Deep residual learning for image recognition](#). In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.
- Luheng He, Mike Lewis, and Luke Zettlemoyer. 2015. [Question-answer driven semantic role labeling: Using natural language to annotate natural language](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 643–653, Lisbon, Portugal. Association for Computational Linguistics.
- Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. [Women also snowboard: Overcoming bias in captioning models](#). In *European Conference on Computer Vision*, pages 793–811.
- Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan C. Russell. 2017. [Localizing moments in video with natural language](#). In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5804–5813. IEEE Computer Society.
- Drew A. Hudson and Christopher D. Manning. 2018. [Compositional attention networks for machine reasoning](#). In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.
- Drew A. Hudson and Christopher D. Manning. 2019. [GQA: A new dataset for real-world visual reasoning and compositional question answering](#). In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6700–6709. Computer Vision Foundation / IEEE.
- Shankar Iyer, Nikhil Dandekar, and Kornel Csernai. [First quora dataset release: Question pairs](#).
- Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik G. Learned-Miller, and Xinlei Chen. 2020. [In defense of grid features for visual question answering](#). In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10264–10273. IEEE.
- Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2017. [CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning](#). In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1988–1997. IEEE Computer Society.
- Kushal Kafle and Christopher Kanan. 2017. [An analysis of visual question answering algorithms](#). In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1983–1991. IEEE Computer Society.
- Divyansh Kaushik, Eduard H. Hovy, and Zachary Chase Lipton. 2020. [Learning the difference that makes A difference with counterfactually-augmented data](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Anna Khoreva, Rodrigo Benenson, Jan Hendrik Hosang, Matthias Hein, and Bernt Schiele. 2017. [Simple does it: Weakly supervised instance and semantic segmentation](#). In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1665–1674. IEEE Computer Society.
- Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. [Bilinear attention networks](#). In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 1571–1581.
- Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2019. [Information maximizing visual question generation](#). In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 2008–2018. Computer Vision Foundation / IEEE.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. [Visual genome: Connecting language and vision using crowdsourced dense image annotations](#). *International journal of computer vision*, 123(1):32–73.



- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. [ALBERT: A lite BERT for self-supervised learning of language representations](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*.
- Patrick Lewis, Ludovic Denoyer, and Sebastian Riedel. 2019a. [Unsupervised question answering by cloze translation](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4896–4910, Florence, Italy. Association for Computational Linguistics.
- Patrick Lewis, Ludovic Denoyer, and Sebastian Riedel. 2019b. [Unsupervised question answering by cloze translation](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4896–4910, Florence, Italy. Association for Computational Linguistics.
- Yikang Li, Nan Duan, Bolei Zhou, Xiao Chu, Wanli Ouyang, Xiaogang Wang, and Ming Zhou. 2018. [Visual question generation as dual task of visual question answering](#). In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6116–6124. IEEE Computer Society.
- Zhongli Li, Wenhui Wang, Li Dong, Furu Wei, and Ke Xu. 2020. [Harvesting and refining question-answer pairs for unsupervised QA](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6719–6728, Online. Association for Computational Linguistics.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.
- Runtao Liu, Chenxi Liu, Yutong Bai, and Alan L. Yuille. 2019. [Clevr-ref+: Diagnosing visual reasoning with referring expressions](#). In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4185–4194. Computer Vision Foundation / IEEE.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. [Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks](#). In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23.
- Mateusz Malinowski and Mario Fritz. 2014. Towards a visual turing challenge. In *Learning Semantics 2014*.
- Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K. Roy-Chowdhury. 2019. [Weakly supervised video moment retrieval from text queries](#). In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 11592–11601. Computer Vision Foundation / IEEE.
- Aditya Mogadala, Marimuthu Kalimuthu, and Dietrich Klakow. 2019. Trends in integration of vision and language research: A survey of tasks, datasets, and methods. *arXiv preprint arXiv:1907.09358*.
- Timothy Niven and Hung-Yu Kao. 2019. [Probing neural network comprehension of natural language arguments](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.
- Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. [Im2text: Describing images using 1 million captioned photographs](#). In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pages 1143–1151.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. [Pytorch: An imperative style, high-performance deep learning library](#). In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. [GloVe: Global vectors for word representation](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.
- Santhosh K. Ramakrishnan, Ambar Pal, Gaurav Sharma, and Anurag Mittal. 2017. [An empirical evaluation of visual question answering for novel objects](#). In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 7312–7321. IEEE Computer Society.



- Arijit Ray, Karan Sikka, Ajay Divakaran, Stefan Lee, and Giedrius Burachas. 2019. [Sunny and dark outside?! improving answer consistency in VQA through entailed question generation](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5860–5865, Hong Kong, China. Association for Computational Linguistics.
- Mengye Ren, Ryan Kiros, and Richard S. Zemel. 2015a. [Exploring models and data for image question answering](#). In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2953–2961.
- Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015b. [Faster R-CNN: towards real-time object detection with region proposal networks](#). In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99.
- Steven Rennie, Etienne Marcheret, Neil Mallinar, David Nahamoo, and Vaibhava Goel. 2020. [Unsupervised adaptation of question answering systems via generative self-training](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1148–1157, Online. Association for Computational Linguistics.
- Marco Tulio Ribeiro, Carlos Guestrin, and Sameer Singh. 2019. [Are red roses red? evaluating consistency of question-answering models](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6174–6184, Florence, Italy. Association for Computational Linguistics.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3).
- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavata, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *AAAI*.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
- Ramprasaath R. Selvaraju, Purva Tendulkar, Devi Parikh, Eric Horvitz, Marco Túlio Ribeiro, Bismira Nushi, and Ece Kamar. 2020. [Squinting at VQA models: Introspecting VQA models with sub-questions](#). In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10000–10008. IEEE.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Improving neural machine translation models with monolingual data](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.
- Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. [Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.
- Robik Shrestha, Kushal Kafle, and Christopher Kanan. 2019. [Answer them all! toward universal visual question answering models](#). In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 10472–10481. Computer Vision Foundation / IEEE.
- Karen Simonyan and Andrew Zisserman. 2015. [Very deep convolutional networks for large-scale image recognition](#). In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- Hyun Oh Song, Ross B. Girshick, Stefanie Jegelka, Julien Mairal, Zaïd Harchaoui, and Trevor Darrell. 2014. [On learning to localize objects with minimal supervision](#). In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 1611–1619. JMLR.org.
- Amos Storkey. 2009. When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning*, pages 3–28.
- Hao Tan and Mohit Bansal. 2019. [LXMERT: Learning cross-modality encoder representations from transformers](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.
- Damien Teney, Ehsan Abbasnejad, and Anton van den Hengel. 2020a. Learning what makes a difference from counterfactual examples and gradient supervision. *arXiv preprint arXiv:2004.09034*.
- Damien Teney, Ehsan Abbasnejad, and Anton van den Hengel. 2020b. Unshuffling data for improved generalization. *arXiv preprint arXiv:2002.11894*.

- Damien Teney and Anton van den Hengel. 2019. [Actively seeking and learning from live data](#). In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 1940–1949. Computer Vision Foundation / IEEE.
- Damien Teney and Anton van den Hengel. 2016. Zero-shot visual question answering. *arXiv preprint arXiv:1611.05546*.
- Damien Teney, Kushal Kafle, Robik Shrestha, Ehsan Abbasnejad, Christopher Kanan, and Anton van den Hengel. 2020c. On the value of out-of-distribution testing: An example of goodhart’s law. *arXiv preprint arXiv:2005.09241*.
- Jörg Tiedemann. 2012. [Parallel data, tools and interfaces in OPUS](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.
- Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. 2019. [Meta-learning to detect rare objects](#). In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 9924–9933. IEEE.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R’emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.
- Jialin Wu and Raymond J. Mooney. 2019. [Self-critical reasoning for robust visual question answering](#). In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8601–8611.
- Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.
- Yiming Xu, Lin Chen, Zhongwei Cheng, Lixin Duan, and Jiebo Luo. 2020. [Open-ended visual question answering by multi-modal domain adaptation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 367–376, Online. Association for Computational Linguistics.
- Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola. 2016. [Stacked attention networks for image question answering](#). In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 21–29. IEEE Computer Society.
- Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. [Deep modular co-attention networks for visual question answering](#). In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6281–6290. Computer Vision Foundation / IEEE.
- Hanwang Zhang, Zawlin Kyaw, Jinyang Yu, and Shih-Fu Chang. 2017. [PPR-FCN: weakly supervised visual relation detection via parallel pairwise R-FCN](#). In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 4243–4251. IEEE Computer Society.
- Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. [Yin and yang: Balancing and answering binary visual questions](#). In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5014–5022. IEEE Computer Society.
- Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. [Learning deep features for discriminative localization](#). In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2921–2929. IEEE Computer Society.

## Appendix

### A Synthesized Samples

Table 7 shows illustrative examples of Q-A pairs procedurally generated from the image caption using template-based method. Table 8 shows the use of two transformations (T): negation and adversarial words (Gokhale et al., 2020b) two generate more sentences. Thus the negation of  $Q$  or substitution of a word in  $Q$  with an adversarial word results in the new question-answer pair  $Q_{new}, A_{new}$ . To increase the linguistic diversity of the questions we use paraphrasing as shown in Table 11.

### B Dataset Analysis

In Table 9, we compare the distribution per answer-type of our synthetically generated samples with the distribution in the VQA-CP-v2 (Agrawal et al.,

Image	Question	Answer
	What are set on the sidewalk outside a veterinary hospital?	bags
	What is the young man holding up in front of his face ?	phone
	What is almost empty on the table	glass
	What drawn carriage with passengers in the city	horse
	What is the color of the table ?	white
	What is the color of the eyes ?	blue
	How many boats anchored by ropes close to shore?	8

Table 7: Examples of template-based data synthesis

T	Image	Q	A	Q <sub>new</sub>	A <sub>new</sub>
Negation		Is this bread?	yes	Is this not bread	no
		What is the color of the woman's shirt?	black	What is not the color of the woman's shirt?	white
		Is there a boy?	no	Is there no boy?	yes
Adversarial		Who is sitting in the boat ?	man	Who is sitting in the dining table ?	can't say
		How big is the plane ?	large	How big is the car ?	size
		How many puppies are on the bed ?	two	How many cats are on the bed?	none

Table 8: The effect of using transformations (T) to create new Q-A pairs

2018) dataset. Since we use our synthetic samples as the pre-training data, and do not use VQA-CP

Category	VQA-CP (%)	Pretraining (%)
Yes/No	41.86	50.18
Number	11.91	8.32
Other	46.23	41.45

Table 9: Distribution of samples by answer-type in our pre-training dataset and the VQA-CP evaluation dataset.

Hyper-Parameters	Model
Batch Size	32-128
Learning Rate	( $1e^{-5}$ , $5e^{-5}$ )
Dropout	0.1
Language Layers	6
Cross-Modality Layer	4 — 12
Optimizer	BertAdam
Warmup	0.1
Max Gradient Norm	5.0
Max Text Length	30
ResNet	50 / 101 / 152
Epochs	10-40

Table 10: Hyper-Parameters for our models

samples for training in our zero-shot setup, this comparison displays the shift between the training (synthetic) and test (human annotated VQA-CP) datasets.

We further analyze this shift, by computing the t-SNE projections of questions using mean-pooled Glove (Pennington et al., 2014) embeddings for our generated questions and observe the overlap with human-authored questions in VQA and GQA (Hudson and Manning, 2019). Figure 6. We observe a marked shift between the question clusters for our procedurally generated questions and human annotated questions from VQA and GQA.

Similarly, we also show the distribution of answers in our dataset in Figure 7. It can be seen that our dataset has a slight imbalance in the proportion of questions with answer “yes” and “no”. Numeric answers 0,1,2,3 are most frequent. Answers about people such as *man*, *woman*, *people*, *person*, *group of people* are also more common in the dataset. The remaining answers have a long-tailed distribution, since there are  $\sim 90k$  unique answers in our dataset compared to  $\sim 3.5k$  in VQA and  $\sim 2k$  in GQA.

## C Training Details

We use the HuggingFace (Wolf et al., 2019) and PyTorch frameworks (Paszke et al., 2019). Hyper-parameters and other training settings are given in Table 10.


Image	Q	A	Q <sub>new</sub>	A <sub>new</sub>
	How is something parked ?	illegally	How's-what's parked?	illegally
	what does something seem to do ?	park	What do you think something seems to be doing?	park
	Where was parked something?	behind a legally parked car	Do you know where something was parked?	behind a legally parked car
	How many cars are visible ?	2	How many cars are we looking at?	2
	Is there two cars parked on the sidewalk on the street ?	Yes	There are two cars parked on the sidewalk, right?	Yes

Table 11: Illustration of using paraphrasing to improve the linguistic variation of our questions and answers.

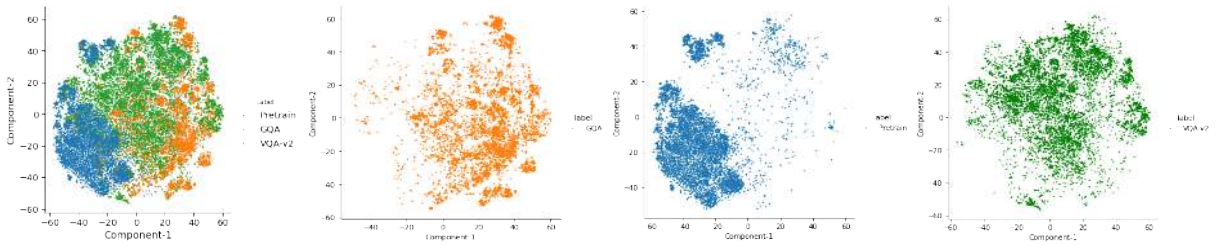


Figure 6: t-SNE projections of GloVe embedding our generated questions, and human-authored VQA-v2 and GQA questions. Blue: our pretraining dataset, Orange: GQA, Green: VQA. L-R: All, GQA, Pretrain, VQA.

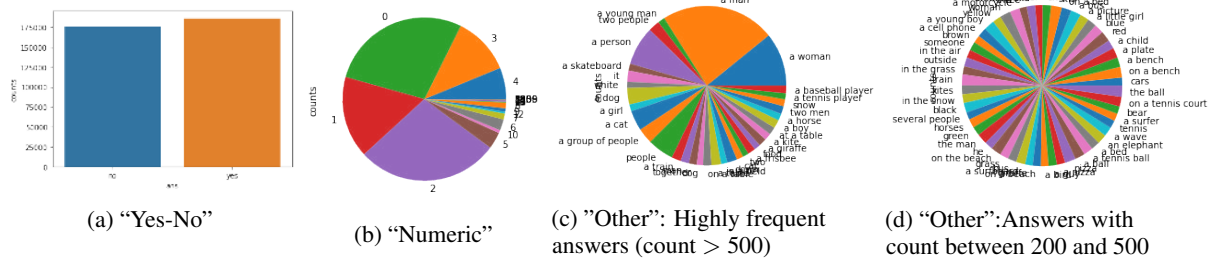


Figure 7: Distribution of most frequent answers in our Pretraining dataset for each answer-type (yes-no, numeric, and other). Please zoom for details.