Visual Question Answering with Annotation-Efficient Zero Shot Learning under Linguistic Domain Shift

Pratyay Banerjee, Tejas Gokhale, Yezhou Yang, Chitta Baral

Arizona State University

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

Abstract

Methodologies for training VQA models assume the availability of datasets with human-annotated *Image-Question-Answer* (I-Q-A) triplets for training. This has led to a heavy reliance and overfitting on datasets and a lack of generalization to new types of questions and scenes. Moreover, these datasets exhibit annotator subjectivity, biases, and errors, along with linguistic priors, which percolate into VQA models trained on such training samples. Captions on the other hand are descriptive and less subjective, and allow us to generate a diverse variety of Q-A pairs. We study whether models can be trained only with images and associated text captions, without any human-annotated Q-A pairs. We train models with procedurally generated Q-A pairs from captions using techniques, such as templates and annotation frameworks like QASRL. Since our Q-A pairs are synthetic, they exhibit a linguistic domain shift from the questions in VQA data and a label-shift in the answer-set, i.e. a zero-shot learning task. As most state-of-the-art VQA models rely on dense and costly object annotations extracted from object detectors, we propose spatial-pyramid image patches as a simple but effective alternative to object bounding boxes, and demonstrate that our method is label-efficient. We benchmark on VQA-v2, GQA, and on VQA-CP which contains a softer version of label shift. Our zero-shot VQA methods surpass prior supervised methods on VQA-CP and approaches state-of-the-art models without object features in fully supervised setting.

1 Introduction

Visual question answering (VQA) has emerged as a crucial task in visual understanding. In fact (Malinowski and Fritz 2014b) posed it as a Turing test (1950), with the goal of building VQA systems indistinguishable from humans. Human-annotated datasets (2014a; 2015; 2018; 2019; 2019) have been used to train and evaluate various VQA models. Unfortunately, heavy reliance on these datasets for training has the unwanted side-effects of bias towards answer styles, question-types (Chao, Hu, and Sha 2018), and spurious correlations with language priors (Agrawal et al. 2018). Similar findings have been reported for natural language datasets (Gururangan et al. 2018; Niven and Kao 2019) and the resulting "Clever Hans Effect" (Kaushik, Hovy, and Lipton 2019). As such, evaluation of VQA models on test-sets

very similar in style to the training samples, is deceptive and inadequate, and not a true measure of robustness.

So what is a way out? One line of work has focused on balancing, de-biasing, and diversifying samples (Goyal et al. 2017; Zhang et al. 2016). However crowdsourcing "unbiased" labels is difficult and costly – it requires a well-designed annotation interface, dedicated and able annotators, and humongous human effort and time (Sakaguchi et al. 2020). The alternative is to avoid the use of annotations, and instead train models in an unsupervised manner by synthesizing training data. These techniques, coined "unsupervised" (Lewis, Denoyer, and Riedel 2019), come with many advantages – first, human bias and subjectivity is reduced; second, 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 methods from CLEVR (Johnson et al. 2017) (which contains artificially rendered images) are also used to generate Q-A pairs for GQA (Hudson and Manning 2019) (which contains real-world and more complex scenes) and also for the referring-expressions task (Liu et al. 2019).

In this work, we train VQA models without using human-annotated Q-A pairs. We utilize image-captioning datasets which provide a multi-perspective concise description of visible objects in an image, and procedurally generate Q-A pairs using a self-supervised mechanism. We train models using this synthetic data and *only evaluate* them on established human-annotated VQA benchmarks: VQA-v2 (Goyal et al. 2017), VQA-CP-v2 (Agrawal et al. 2018), and GQA (Hudson and Manning 2019).

Why Captions? Image captioning, like VQA, is a cornerstone of current vision-and-language research, and datasets such as MS-COCO (Lin et al. 2014) contain captions that describe visual entities such as objects and actions in images of common objects and everyday scenes. During the construction of the dataset (Chen et al. 2015), human caption writers were instructed to refrain from describing things that have happened in the past or in future, and "what a person might say". One 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 easily answer" and "interesting" questions that may require "commonsense". A different set of annotators provided answers to these ques-

tions and were allowed to "speculate" an answer that "most people would agree on" or to simply provide a best guess. For instance in Figure 2, the first VQA-v2 question is an example of linguistic bias since most cars have four doors, and the second question is subjective and so has multiple contradicting answers from different annotators. Similarly the first GQA question is ambiguous and could refer to either the skier or the photographer. It has been shown multiple answers may exist for questions in common VQA datasets due to perceived difficulty of the question, ambiguity, and subjectivity (Bhattacharya, Li, and Gurari 2019). Thus the very nature of the data-collection procedure and instructions for VQA brings in human subjectivity and linguistic bias as compared to image-captioning annotations which are designed to be simple, precise, and non-speculative. Moreover, procedurally generating Q-A pairs from captions one can create a diverse variety of questions that need deep image understanding. These are our motivations for using captions to synthesize Q-A pairs.

For the creation of Q-A pairs from image captions, we use template-based methods similar to (Ren, Kiros, and Zemel 2015; Gokhale et al. 2020), along with paraphrasing and back-translation (Sennrich, Haddow, and Birch 2016) for lingustic variation. We also synthesize questions about image semantics using the QA-SRL (He, Lewis, and Zettlemoyer 2015) approach. Since our Q-A pairs are created synthetically, there exists a domain shift as well as label (answer) shift from evaluation datasets such as VQA-v2 and GQA as shown in Figure 2, making this task zero-shot.

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 annotations needed to extract object features using Faster R-CNN (Ren et al. 2015), we propose spatial pyramids of image patches as a simple, effective, and annotation-efficient alternative. To the best of our knowledge, this is the first work on the unsupervised visual question answering, with the following additional contributions:

- We introduce the self-supervised data synthesis framework for creating Q-A pairs from captions, which include multi-word answer phrases.
- We propose pre-training tasks that use spatial pyramids of image-patches instead of object bounding-boxes, further making our method label-efficient.
- We perform extensive experiments and analyses under zero-shot and fully-supervised settings, and establish benchmarks on VQA-v2, VQA-CP, and GQA.
- Our model achieves state of the art accuracy on the zero-shot VQA task and thus serves as a strong baseline for future work on zero-shot VQA.

2 The Many Faces of Generalization in VQA

Preliminaries: Given an input question Q about an image I, the goal for visual question answering is to provide an answer A. Thus for a VQA model, the (Q, I) pair belongs to the domain, while the answer belongs to the label-space.

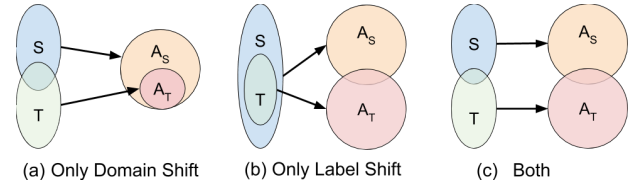Let S be the space of source questions available for train-

(a) Only Domain Shift (b) Only Label Shift (c) Both

Figure 1: Aspects of generalization: domain shift implies that $S \cap T \neq T$, while label shift implies that $A_S \cap A_T \neq A_T$.

ing with the corresponding answer-set $\mathcal{A}_S$, and T that of target questions with answer-set $\mathcal{A}_T$. The aim of a VQA model is to generalize to T. Let $\mathcal{I}$ be the set of images and $\mathcal{Q}_S$ and $\mathcal{Q}_T$ be the set of questions in source and target datasets, both of which are split into sets of training and test samples. Let $\mathcal{D}_S, \mathcal{D}_T$ denote the source and target datasets, given by:

$$\mathcal{D}_S^{train} = \{(I, Q) | Q \in \mathcal{Q}_S^{train}, I \in \mathcal{I}^{train}\}, \quad (1)$$

$$\mathcal{D}_S^{test} = \{(I, Q) | Q \in \mathcal{Q}_S^{test}, I \in \mathcal{I}^{test}\}, \quad (2)$$

$$\mathcal{D}_T^{test} = \{(I, Q) | Q \in \mathcal{Q}_T^{test}, I \in \mathcal{I}^{test}\}. \quad (3)$$

$\mathcal{I}^{train}, \mathcal{I}^{test}$ are independent and identically distributed and $\mathcal{I}^{train} \cap \mathcal{I}^{test} = \emptyset$. Then two aspects of generalization can be defined as shown in Figure 1, *domain shift* i.e. generalization to a new domain of inputs, characterized by $S \cap T \neq T$; and *label-shift*, i.e. generalization to novel answers, characterized by $A_S \cap A_T \neq A_T$.

Related Work: Work in VQA has recently attracted attention from various unique (but scattered) points of view such as robustness, reduction of biases and spurious correlations, and transfer to other question-types. Unfortunately, such work has not been formulated in the conventional language of domain generalization. We review related work and unite it under the umbrella of the formalism of Figure 1.

Performance under **domain shift** has been evaluated for new domains of test questions with unseen words (Teney and Hengel 2016), unseen objects (Ramakrishnan et al. 2017), novel compositions (Johnson et al. 2017; Agrawal et al. 2017) and logical connectives (Gokhale et al. 2020). Adaptation to different datasets with varying linguistic styles (Chao, Hu, and Sha 2018; Xu et al. 2019; Shrestha, Kafle, and Kanan 2019) and different reasoning capabilities (Kafle and Kanan 2017) has also been studied. Other work seeks to answer target questions that are sub-questions (Selvaraju et al. 2020), or are implied (Ribeiro, Guestrin, and Singh 2019) or entailed (Ray et al. 2019) by source questions.

Label shift or Prior Probability Shift (Storkey 2009) has been implicitly hinted at in the VQA-CP challenge (Agrawal et al. 2018), where the conditional probabilities of answers given the question, $P(A|Q)$, for the train and test splits are designed to be distant. The ability to perform a task amidst label-shift is typically termed as **zero-shot learning** (Lampert, Nickisch, and Harmeling 2009; Farhadi et al. 2009; Palatucci et al. 2009). In the VQA-v2 dataset (Goyal et al. 2017), questions are created for each image $i \in \mathcal{I}$ by human workers, and the dataset is balanced with respect to $P(A|Q)$. This is an example of neither domain shift nor label shift.

Captions		Question	Answer(Confidence)
- A car that seems to be parked illegally behind a legally parked car - A couple of cars parked in a busy street sidewalk - Cars try to maneuver into parking spaces along a densely packed street. - two cars parked on the sidewalk on the street	**VQA-v2**	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)
	Synthetic (Ours)	1. How is something parked ? 2. Is there a truck ? 3. Is it a couple of cars parked in a busy street sidewalk? 4. Where does something maneuver?	Illegally (1.0) No (1.0) Yes (1.0) Into Parking Spaces (1.0)
- A man in skies is coming up the hill - A skier is passing a competition race marker - A man takes a picture of a skier - A cross-country skier is competing at night in snow	**GQA**	1. Is the man on the left or on the right ? 2. Who is wearing the jersey ?	Right (1.0) Man (1.0)
	Synthetic (Ours)	1. What is someone passing ? 2. When is someone competing ? 3. Who is coming ? 4. Is that a man in skateboard coming up the hill ? 5. Where is someone coming?	A competition race marker (1.0) At night (1.0) A man in skiis (1.0) No Up the hill (1.0)

More examples can be found in the Appendix.

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

Our Problem Statement: Our work deals with learning VQA using only images with associated captions, without any labeled Q-A pairs. Consider an image captioning dataset $\mathcal{D}_C$ with captions C for each image in $\mathcal{I}^{train}$ as shown in Figure 2. Using these captions, we wish to answer questions about test-images in $\mathcal{I}^{test}$ such that none of these images were previously observed ($\mathcal{I}^{train} \cap I^{test} = \emptyset$). For this, we create synthetic questions $\mathcal{Q}_S^{train}$ and answers $\mathcal{A}_S^{train}$ for training. As such, a linguistic domain shift exists between these synthetic source questions and human-annotated target questions from datasets such as VQA-v2 and GQA. Similarly, owing to the automated nature of our Q-A pair generation, a label-shift is also observed. In this paper, for the first time, we address the harder and unexplored problem of zero-shot VQA trained on procedurally generated samples exhibiting both domain-shift and label-shift.

3 Self-Supervised Q-A Synthesis Framework

In this section, we detail our framework for procedurally generating Q-A pairs using captions as input. Captions C and object-words O are used as input. These object words are estimated from the caption by using simple heuristics such as, extracting noun-phrases and using numerical quantifiers in the caption as soft approximations of the cardinality of objects, using Spacy (Honnibal and Montani 2017). For example, if a caption is "There are four apples placed on a basket.", we extract {"apples", "basket"} as the objects, and {4, 1} as their respective counts. If object-words are available explicitly, we used them as is.

3.1 Question Generation

Question generation in itself is a complex domain and several studies are dedicated to it (2017; 2019). We approach it conservatively, using template-based methods and QA-SRL-based semantic role labeling for question generation, and with paraphrasing and backtranslation for improving the linguistic diversity of template-based questions. Questions are categorized based on their answer types; *Yes-No*, *Number*, *Color*, *Location*, *Object* and *Phrases*.

**Template-based:** To create *Yes-No* questions, the caption is processed as follows: first modal verbs are removed and then, a randomly chosen question prefix such as *"is there", "is this", "does this look like"* 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", for which the answer is "Yes". To create the corresponding question with answer "No", we use either a 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 not in the image, but similar to objects in the image. To compute similarity, we use Glove (2014) vector embeddings.

For *Object*, *Number*, *Location*, and *Color* questions, we follow the procedure similar to COCO-QA (2015). To create "what" questions for the *Object* type, we extract objects from captions as noun phrases, replace them with "what", and rephrase the question such that it starts with "what". The rephrasing is done by first splitting long sentences to shorter ones, converting indefinite determiners to definite, replacing potential answer options with "what". We follow a similar procedure for *Number* questions; we extract numeric quantifiers of noun phrases, and ask "how many" and "what is the count" questions. *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 (2015) was proposed as a paradigm to use natural language to annotate NLP data, by using Q-A pairs to specify textual arguments and their roles. Consider the caption *"A girl in a red shirt holding a skateboard 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 *"when", "what", "where", and "who"* questions, we obtain the following Q-A pairs belonging to the *Phrases* category:

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

QA-SRL questions are short and use generic descriptors and pronouns such as *something* and *someone* instead of
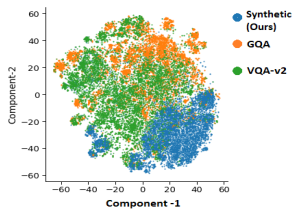
	Template-based	Paraphrase & Back-translate	QA-SRL	VQA-v2	GQA	VQA-CP
# of Questions	600K	400K	2.5M	438K / 214K	943K / 132K	245K / 220K
# of Answers	5K	5K	90K	3.5K	1878	3.5K
Mean Question Length	7.9	8.1	4.8	6.4	10.6	6.4
Mean Answer Length	1.4	1.4	6.3	1.1	1.3	1.1
Image Source	COCO	COCO	COCO	COCO	COCO,Genome,Flickr	COCO
Image Counts	204K	204K	204K	204K	113K	204K

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

elaborate references, while the expected answer phrases are longer and descriptive as shown above. Thus to answer these, greater semantic understanding of the image is required. For captions that do not contain verbs that can be labelled with semantic roles, we skip QA-SRL and only use template-based methods for question generation.

Paraphrasing and Back-Translation: In order to increase the linguistic variation in the questions, we apply two natural language data augmentation techniques, paraphrasing and back-translation. To paraphrase questions, we train a T5 (Raffel et al. 2019) text generation model on the Quora Question Pairs Corpus (2017). 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 re-translate the question back to English. For example, *"Is the girl who is to the left of the sailboats wearing a backpack?"* is translated to *''La chica que está a la izquierda de los veleros lleva mochila?"* in Español, and back-translated to *"Does the girl to the left of the sailboats carry a backpack?"*.

3.2 Comparitive Analysis with VQA-v2 and GQA

Answers to QA-SRL questions are more descriptive with use of adjectives, adverbs, determiners, and quantifiers, compared to current VQA benchmarks, which typically contain one-word answers, as seen in Figure 2. Similarly, 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. We observe there are around 200 answers that 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.

The style of some of our synthetic questions such as counting questions, object presence/absence questions created by template-based question generation, is also found in VQA and GQA. On the other hand, QA-SRL questions require semantic understanding of the actions (verb) depicted in the image, which are rare in VQA and GQA. We quantify this by plotting the t-SNE components of document vector embeddings of the questions from VQA, GQA and our synthetic data, in Figure 3. We can observe that the human-annotated questions from VQA and GQA have a significant overlap, whereas our synthetic dataset questions are a distinct cluster. Learning to perform visual question answering from disparate linguistic styles, and evaluating on conventional benchmarks is the challenge we address in this paper.

4 Method

In this section, we describe our visual question answering model. Recent approaches, such as LXMERT (Tan and Bansal 2019), VilBERT (Lu et al. 2019), and UNITER (Chen et al. 2019) all use deep transformer-encoder architectures and pre-train using a combination of multiple image captioning and VQA datasets such as Conceptual Captions (Sharma et al. 2018), SBU Captions (Ordonez, Kulkarni, and Berg 2011), Visual Genome (Krishna et al. 2017), and MSCOCO (Lin et al. 2014). Training on such a huge collection of data is resource-intensive and hence we train our models only on MSCOCO captions. Moreover MSCOCO captions are of good quality, less noisy, and provide multi-perspective descriptions for each image.

4.1 Spatial Pyramid Patches

"Bottom-Up" object region features (Anderson et al. 2018) extracted from Faster R-CNN have become the de-facto image features used in state-of-the-art VQA models, which do not use the entire image, but only the features of the detected objects as inputs for QA. Although object features are discriminative, dense annotations are required for training and additional large deep networks for extraction. Moreover, object detection can be imperfect. For example, 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. Similarly, object detection is not reliable for detecting rare objects (Wang, Ramanan, and Hebert 2019). This is a problematic bottle-neck for VQA performance on 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 an image classification task on ImageNet (Russakovsky et al. 2015) have been previously used for VQA (2017). ImageNet mostly contains iconic (single-object) images, making the feature extraction restrictive, as VQA is often about non-iconic images, with questions about relations between multiple objects. Inspired by Spatial Pyramid Matching (2006) for image classification, we propose *spatial pyramid patch features* to represent the input image into a sequence of features at different scales. For an input image I, we use a ResNet (pretrained on ImageNet) to extract features from a set of image patches $\{I_{k_1}, \ldots, I_{k_n}\}$, where I_{k_i} is the image divided into a $k_i \times k_i$ grid of patches. Larger patches encode global features and relations, while smaller patches encode local and low-level features.
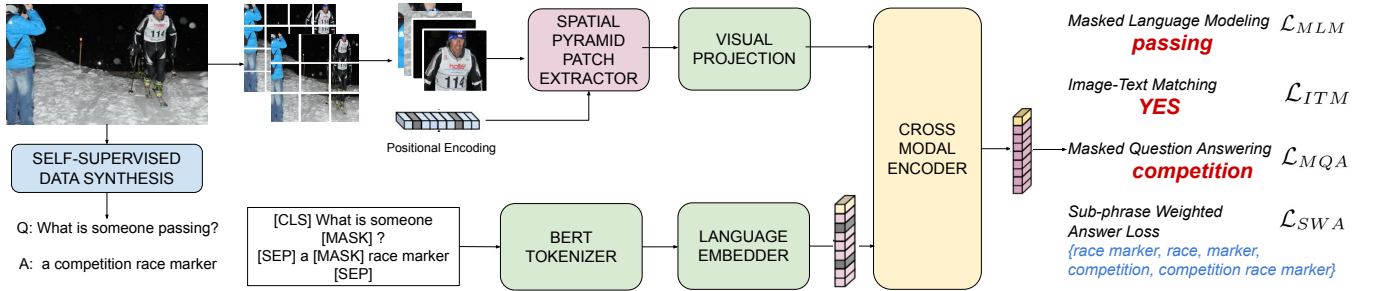
Figure 4: Our model architecture makes the use of spatial pyramids of image patches as inputs to the Encoder, which is trained for three pre-training tasks as shown.

4.2 Encoder

Our Encoder model is similar to the UNITER single-stream transformer, where the sequence of word tokens $w = \{w_1, ..., w_T\}$ and the sequence of image patch features $v = \{v_1, ..., v_K\}$ are taken as input. We tokenize the text using a WordPieces (Wu et al. 2016) tokenizer similar to BERT (Devlin et al. 2018), and embed the text tokens through a text-embedder. 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, and a residual connection and layer normalization is added after cross-modal attention layer similar to Vaswani et al..

4.3 Pre-training Tasks

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", and is trained to predict this masked token. Without the image, there can be multiple plausible choices, 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).

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) For each image, we use the five captions provided by MS-COCO as positive samples. 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 (w_1), and feed it to a feed-forward layer followed by a softmax. For MLM and MQA we feed corresponding token representations to a different feed-forward layer. We train the model using standard cross-entropy loss for all three tasks.

4.4 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 $W_{sub} = WordCount(sub)/WordCount(ans)$. 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 given the input question, the model predicts a weighted distribution y_{wa} over the answer vocabulary.

$$\mathcal{L}_{SWA} = \mathcal{L}_{BCE}(\sigma(z^{[CLS]}), y_{wa}). \qquad (4)$$

Finally the full loss function with α and β as loss scale factors is:

$$\mathcal{L} = \mathcal{L}_{MLM} + \mathcal{L}_{MQA} + \alpha\mathcal{L}_{ITM} + \beta\mathcal{L}_{SWA}. \qquad (5)$$

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* (trained only on procedurally generated samples), and *fully-supervised* (where we finetune our model using the associated train annotations) settings. We report exact-match accuracies as our metrics for evaluation.

Training Our Encoder has 8 cross-modal layers with a hidden dimension of 768. 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

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 Hengel 2019)	46.0	58.2	29.5	44.3
Unshuffling (2020)	42.4	47.7	14.4	47.3
UpDn+CE+GS (2020)	46.8	64.5	15.4	45.9
LXMERT (2019)	46.2	42.8	18.9	**55.5**
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	**47.3**	<u>73.4</u>	**39.8**	<u>35.6</u>

Table 1: Unsupervised accuracy on VQA-CP-v2 test set. All baselines are supervised methods trained on the train split. [2]

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
ViLBERT (2019)	70.5	*	*	*
LXMERT (2019)	72.5	**88.2**	**54.2**	**63.1**
UNITER (2019)	**72.7**	*	*	*
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 + 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[2]. FSL models are pretrained on synthetic samples, and further finetuned on VQA-v2 train split. *not available

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)	**60.3**	**77.8**	**45.0**
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 + Patches + UpDown	46.4	64.3	31.4
FSL + Patches + Encoder	55.2	73.6	38.8

Table 3: GQA Validation split accuracies.[2]

Zero-shot Question Answering   Tables 1, 2 and 3 summarize our results on the three benchmark datasets respectively. We can observe that our method outperforms specially designed supervised methodsfor bias removal in VQA-CP. Our procedurally generated Q-A pairs improve performance for both UpDown and Encoder models, showing the method to be effective, and that the improvements are model-agnostic. Our Encoder model further improves the performance by 5.5% over the UpDown baseline. In the zero-shot setting, compared to object-features, our Spatial Image Patch features perform equally well on VQA, and are better on VQA-CP, and are also more annotation efficient. In GQA, the zero-shot performance is not as competetive when compared to our performance on VQA and VQA-CP. We attribute this to the need for understanding spatial relationships answer GQA questions. Such questions are infrequent in our synthetic training data since human-annotated captions do not contain detailed spatial relationships among objects. The development of self-supervised techniques to perform spatial reasoning is an interesting future direction for research.

Fully Supervised Question Answering   In the fully supervised setting, the performance of our methods approaches SOTA methods. However, our methods are significantly annotation-efficient as we only adopt COCO captions without dense object annotations during pre-training or training. In GQA, the Encoder model performs on par with MAC (2018) and BAN (2018), which unlike us, use object relationship annotations. This suggests that pyramidal features and the cross-modal transformer encoder layers can learn spatial relationships between image regions.

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 overlapping spatial im-

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. All our models are trained using 4 Nvidia V100 16 GB GPUs.

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 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.

6   Results

In this section, we discuss our results and outcomes from analyses. ZSL refers to zero-shot setting and FSL refers to our models further finetuned on the respective train split.

---

[2]In all tables <u>underline</u> implies unsupervised best, and **bold** implies overall best. Baselines are trained on VQA/VQA-CP/GQA training data and our models on synthetic self-supervised data.

	Datasets	Template	Template + P & B	QASRL	All
Updn	VQA-v2	26.2	28.5	31.1	41.4
	VQA-CP	25.7	27.1	33.8	40.2
	GQA	11.6	14.8	18.9	31.1
Encoder	VQA-v2	32.5	34.8	40.3	**47.1**
	VQA-CP	31.2	33.6	39.8	**46.8**
	GQA	18.5	23.6	21.4	**33.7**

Table 4: Effect of different training data sources on ZSL validation accuracy. P&B Paraphrasing and Back-translation.

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

Table 5: Effect of the number of spatial patches on ZSL validation accuracies with UpDn and our Encoder. {3,5} implies division of the image into a 3x3 and 5x5 grid of patches.

Datasets	SWA	MLM+ SWA	MQA+ SWA	MLM+MQA +SWA	MLM+IT SWA	All
VQA-v2	39.1	42.4	42.0	45.6	44.7	**46.2**
VQA-CP	38.3	41.5	41.2	44.9	43.6	**45.4**
GQA	25.4	27.8	26.6	29.7	28.9	**31.2**

Table 6: Effect of different Pre-training tasks on the ZSL validation accuracies for the Encoder model.
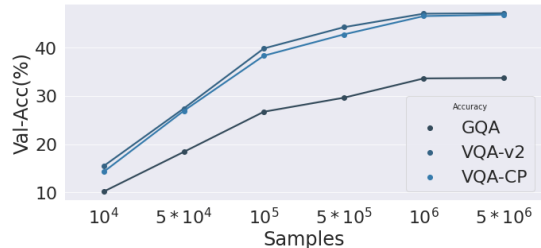

Figure 5: Learning Curve showing validation accuracy vs. the number of synthetically generated training samples.

age patches (i.e. decreasing the patch size). It can be observed in Table 5 that an optima exists at grid-size of $7x7$ after which the addition of smaller patches is detrimental. Similarly, only using patches of large size does not allow models to focus on specific regions of the image. Thus a trade-off exists between global context and region-specific features. We observe a minor improvement of 0.01-0.3% by extracting features from ResNet-101 compared to ResNet-50. Removing visual position embeddings has a significant effect on performance, with a drop of 4.6-8% on average, in both ZSL and FSL settings for VQA and GQA.

Effect of different Pre-training Tasks Table 6 shows the effect of different pretraining tasks on the downstream zeroshot VQA task. We need the SWA task, as it is used to perform the zeroshot 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 Train set Figure 1 shows the learning curve of our Encoder model for the zeroshot 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, as compared to training on the entire set.

Error Analysis Our ZSL method is pretrained on longer phrases and hence tends to generate answers with more details, such as "red car" instead of "car". Although the SWA loss mitigates this to an extent, by creating a distribution over the shorter phrases, the bias is not completely 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 our detailed predictions are also plausible answers.

This not only shows the relevance of learning from captions, but also quantifies the bias towards short "true" answers in human-annotated benchmarks, demonstrating the need for better evaluation metrics that do not penalize VQA systems for producing descriptive accurate answers.

In the fully supervised setting, we either finetune our pre-trained QA classifier with the SWA Loss, or train a separate feedforward layer for the task. The pre-trained QA classifier continues to predict longer phrases as answers, leading to a drop in accuracy. The feedforward layer (trained from scratch) 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.

7 Discussion and Conclusion

Prior work (Chen et al. 2019) has effectively demonstrated that the use of object bounding-boxes and region features leads to significant improvements on downstream tasks such as captioning and VQA. But little effort has been dedicated towards developing alternative methods that can approach similar performance without relying on dense annotations. We argue that annotation-efficiency, self-supervised learning, and data synthesis techniques could be the pathway for the V& L community towards a "post-dataset era[1]". In this work, we take a step towards that goal. We present a framework for procedural synthesis of Q-A pairs, and introduce the new task of zero-shot visual question answering, 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 analysis demonstrates problems with existing VQA evaluation metrics. To mitigate this, we introduce the subphrase weighted answer loss. Our method surpasses previous supervised methods on VQA-CP.

---

[1]A. Efros, *Imagining a post-dataset era*, ICML'20 Invited Talk.

Ethical Considerations

Captions, although also collected from human annotators, have been shown to have a lesser degree of subjectivity, ambiguity, and linguistic biases than VQA datasets, due to the design of annotation prompts that limit the introduction of these biases. Our work has demonstrated that procedurally generated annotations can help mitigate linguistic priors in VQA models (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.

References

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

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

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *CVPR*.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433.

Bhattacharya, N.; Li, Q.; and Gurari, D. 2019. Why Does a Visual Question Have Different Answers? In *Proceedings of the IEEE International Conference on Computer Vision*.

Cadene, R.; Dancette, C.; Ben younes, H.; Cord, M.; and Parikh, D. 2019. RUBi: Reducing Unimodal Biases for Visual Question Answering. In *Advances in Neural Information Processing Systems 32*, 841–852. URL http://papers.nips.cc/paper/8371-rubi-reducing-unimodal-biases-for-visual-question-answering.pdf.

Chao, W.-L.; Hu, H.; and Sha, F. 2018. Cross-dataset adaptation for visual question answering. In *CVPR*, 5716–5725.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325* .

Chen, Y.-C.; Li, L.; Yu, L.; Kholy, A. E.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2019. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740* .

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Du, X.; Shao, J.; and Cardie, C. 2017. Learning to ask: Neural question generation for reading comprehension. *arXiv preprint arXiv:1705.00106* .

Farhadi, A.; Endres, I.; Hoiem, D.; and Forsyth, D. 2009. Describing objects by their attributes. In *CVPR*, 1778–1785.

FitzGerald, N.; Michael, J.; He, L.; and Zettlemoyer, L. 2018. Large-Scale QA-SRL Parsing. In *56th Annual Meeting of the ACL*.

Gokhale, T.; Banerjee, P.; Baral, C.; and Yang, Y. 2020. VQA-LOL: Visual question answering under the lens of logic. In *European Conference on Computer Vision (ECCV)*.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *CVPR*, 6904–6913.

Grand, G.; and Belinkov, Y. 2019. Adversarial Regularization for Visual Question Answering: Strengths, Shortcomings, and Side Effects. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, 1–13.

Gurari, D.; Li, Q.; Stangl, A. J.; Guo, A.; Lin, C.; Grauman, K.; Luo, J.; and Bigham, J. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*.

Gururangan, S.; Swayamdipta, S.; Levy, O.; Schwartz, R.; Bowman, S. R.; and Smith, N. A. 2018. Annotation Artifacts in Natural Language Inference Data. In *NAACL-HLT (2)*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

He, L.; Lewis, M.; and Zettlemoyer, L. 2015. Question-answer driven semantic role labeling: Using natural language to annotate natural language. In *Proceedings of 2015 the conference on EMNLP*, 643–653.

Hendricks, L. A.; Burns, K.; Saenko, K.; Darrell, T.; and Rohrbach, A. 2018. Women also snowboard: Overcoming bias in captioning models. In *European Conference on Computer Vision*, 793–811.

Honnibal, M.; and Montani, I. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing .

Hudson, D. A.; and Manning, C. D. 2018. Compositional Attention Networks for Machine Reasoning. In *International Conference on Learning Representations*.

Hudson, D. A.; and Manning, C. D. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 6700–6709.

Iyer, S.; Dandekar, N.; and Csernai, K. 2017. First Quora Dataset Release: Question Pairs. URL https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs.

Johnson, J.; Hariharan, B.; van der Maaten, L.; Fei-Fei, L.; Lawrence Zitnick, C.; and Girshick, R. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*.

Kafle, K.; and Kanan, C. 2017. An analysis of visual question answering algorithms. In *Proceedings of the IEEE International Conference on Computer Vision*, 1965–1973.

Kaushik, D.; Hovy, E.; and Lipton, Z. 2019. Learning The Difference That Makes A Difference With Counterfactually-Augmented Data. In *International Conference on Learning Representations*.

Kim, J.-H.; Jun, J.; and Zhang, B.-T. 2018. Bilinear attention networks. In *Advances in Neural Information Processing Systems*, 1564–1574.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision* 123(1): 32–73.

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*.

Lazebnik, S.; Schmid, C.; and Ponce, J. 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*.

Lewis, P.; Denoyer, L.; and Riedel, S. 2019. Unsupervised Question Answering by Cloze Translation. In *Proceedings of the 57th Annual Meeting of the ACL*, 4896–4910.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Liu, R.; Liu, C.; Bai, Y.; and Yuille, A. L. 2019. Clevr-ref+: Diagnosing visual reasoning with referring expressions. In *CVPR*, 4185–4194.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Vilbert: Pre-training task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, 13–23.

Malinowski, M.; and Fritz, M. 2014a. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in neural information processing systems*, 1682–1690.

Malinowski, M.; and Fritz, M. 2014b. Towards a Visual Turing Challenge. In *Learning Semantics 2014*.

Niven, T.; and Kao, H.-Y. 2019. Probing Neural Network Comprehension of Natural Language Arguments. In *57th Annual Meeting of the ACL*, 4658–4664.

Ordonez, V.; Kulkarni, G.; and Berg, T. L. 2011. Im2text: Describing images using 1 million captioned photographs. In *Advances in neural information processing systems*.

Palatucci, M.; Pomerleau, D.; Hinton, G. E.; and Mitchell, T. M. 2009. Zero-shot learning with semantic output codes. In *Advances in neural information processing systems*, 1410–1418.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on EMNLP*, 1532–1543.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683* .

Ramakrishnan, S. K.; Pal, A.; Sharma, G.; and Mittal, A. 2017. An empirical evaluation of visual question answering for novel objects. In *CVPR*, 4392–4401.

Ray, A.; Sikka, K.; Divakaran, A.; Lee, S.; and Burachas, G. 2019. Sunny and Dark Outside?! Improving Answer Consistency in VQA through Entailed Question Generation. In *Proceedings of the 2019 Conference on EMNLP*.

Ren, M.; Kiros, R.; and Zemel, R. 2015. Exploring models and data for image question answering. In *Advances in neural information processing systems*, 2953–2961.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 91–99.

Ribeiro, M. T.; Guestrin, C.; and Singh, S. 2019. Are Red Roses Red? Evaluating Consistency of Question-Answering Models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6174–6184.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115(3).

Sakaguchi, K.; Bras, R. L.; Bhagavatula, C.; and Choi, Y. 2020. WINOGRANDE: An Adversarial Winograd Schema Challenge at Scale. In *AAAI*.

Selvaraju, R. R.; Tendulkar, P.; Parikh, D.; Horvitz, E.; Ribeiro, M. T.; Nushi, B.; and Kamar, E. 2020. SQuINTing at VQA Models: Introspecting VQA Models With Sub-Questions. In *CVPR*, 10003–10011.

Sennrich, R.; Haddow, B.; and Birch, A. 2016. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the ACL*.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the ACL*, 2556–2565.

Shrestha, R.; Kafle, K.; and Kanan, C. 2019. Answer them all! toward universal visual question answering models. In *CVPR*, 10472–10481.

Storkey, A. 2009. When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning* 3–28.

Tan, H.; and Bansal, M. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *EMNLP 2019*.

Teney, D.; Abbasnedjad, E.; and Hengel, A. v. d. 2020. Learning what makes a difference from counterfactual examples and gradient supervision. *arXiv preprint arXiv:2004.09034* .

Teney, D.; Abbasnejad, E.; and Hengel, A. v. d. 2020. Un-shuffling Data for Improved Generalization. *arXiv preprint arXiv:2002.11894* .

Teney, D.; and Hengel, A. v. d. 2016. Zero-shot visual question answering. *arXiv preprint arXiv:1611.05546* .

Teney, D.; and Hengel, A. v. d. 2019. Actively seeking and learning from live data. In *CVPR*, 1940–1949.

Tiedemann, J. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.

Turing, A. 1950. Computing Machinery and Intelligence. *Mind* 59(236): 433–460.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wang, Y.-X.; Ramanan, D.; and Hebert, M. 2019. Meta-learning to detect rare objects. In *Proceedings of the IEEE International Conference on Computer Vision*, 9925–9934.

Wu, Y.; Schuster, M.; Chen, Z.; Le, Q. V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144* .

Xu, Y.; Chen, L.; Cheng, Z.; Duan, L.; and Luo, J. 2019. Open-Ended Visual Question Answering by Multi-Modal Domain Adaptation. *arXiv preprint arXiv:1911.04058* .

Yang, Z.; He, X.; Gao, J.; Deng, L.; and Smola, A. 2016. Stacked attention networks for image question answering. In *CVPR*, 21–29.

Yu, Z.; Yu, J.; Cui, Y.; Tao, D.; and Tian, Q. 2019. Deep modular co-attention networks for visual question answering. In *CVPR*.

Zellers, R.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. From Recognition to Cognition: Visual Commonsense Reasoning. In *CVPR*.

Zhang, P.; Goyal, Y.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2016. Yin and yang: Balancing and answering binary visual questions. In *CVPR*.