VQA-LOL: Visual Question Answering under the Lens of Logic

Tejas Gokhale[*][0000−0002−5593−2804], Pratyay Banerjee [*][0000−0001−5634−410X],
Chitta Baral[0000−0002−7549−723X], and Yezhou Yang[0000−0003−0126−8976]

Arizona State University, United States
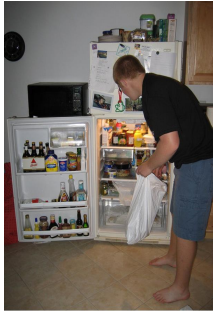{tgokhale, pbanerj6, chitta, yz.yang}@asu.edu

Fig. 1: State-of-the-art models answer questions from the VQA dataset (Q_1, Q_2) correctly, but struggle when asked a logical composition including negation, conjunction, disjunction, and antonyms. We develop a model that improves on this metric substantially, while retaining VQA performance.

Abstract. Logical connectives and their implications on the meaning of a natural language sentence are a fundamental aspect of understanding. In this paper, we investigate whether visual question answering (VQA) systems trained to answer a question about an image, are able to answer the logical composition of multiple such questions. When put under this *Lens of Logic*, state-of-the-art VQA models have difficulty in correctly answering these logically composed questions. We construct an augmentation of the VQA dataset as a benchmark, with questions containing logical compositions and linguistic transformations (negation, disjunction, conjunction, and antonyms). We propose our Lens of Logic (LOL) model which uses question-attention and logic-attention to understand logical connectives in the question, and a novel Fréchet-Compatibility Loss, which ensures that the answers of the component questions and the composed question are consistent with the inferred logical operation. Our model shows substantial improvement in learning logical compositions while retaining performance on VQA. We suggest this work as a move towards robustness by embedding logical connectives in visual understanding.

Keywords: Visual Question Answering, Logical Robustness

---

[*] Equal Contribution

1   Introduction

Theories about logic in human understanding have a long history. In modern times, Piaget and Fodor [34] studied the representation of logical hypotheses in the human mind. George Boole [7] formalized conjunction, disjunction, and negation into an "algebra of thought" as a way to improve, systemize, and mathematize Aristotle's Logic [12]. Horn regarded negation to be a fundamental and defining characteristic of human communication [19], following the traditions of Sankara [35], Spinoza [42], and Hegel [18]. Recent studies [11] have suggested that infants can formulate intuitive and stable logical structures to interpret dynamic scenes and to entertain and rationally modify hypotheses about the scenes. As such we argue that understanding logical structures in questions, is a fundamental requirement for any question-answering system.

If a question can be put at all, then it can be answered. [44]

In the above proposition, Wittgenstein linked the process of asking a question with the existence of an answer. While we do not comment on the existence of an answer, we suggest the following softer proposition -

If questions $Q_1 \ldots Q_n$ can be answered, then so should all composite questions created from $Q_1 \ldots Q_n$

Visual question answering (VQA) [3] is an intuitive, yet challenging task that lies at a crucial intersection of vision and language. Given an image and a question about it, the goal of a VQA system is to provide a free-form or open-ended answer. Consider the image in Figure 1 which shows a person in front of an open fridge. When asked the questions Q_1 (*Is there beer?*) and Q_2 (*Is the man wearing shoes?*) independently, the state-of-the-art model LXMERT [43] answers both correctly. However when we insert a negation in Q_2 (*Is the man not wearing shoes?*) or for a conjunction of two questions $\neg Q_2 \wedge Q1$ (*Is the man not wearing shoes and is there beer?*), the system makes wrong predictions. Our motivation is to reliably answer such logically composed questions. In this paper, we analyze VQA systems under this *Lens of Logic (LOL)* and develop a model that can answer such questions reflecting human logical inference. We offer our work as the first investigation into the logical structure of questions in visual question-answering and provide a solution that *learns* to interpret logical connectives in questions.

The first question is: can models pre-trained on the VQA dataset answer logically composed questions? It turns out that these models are unable to do so, as illustrated in Figure 1 and Table 2. An obvious next experiment is to *split the question* into its component questions, predict the answer to each, and combine the answers logically. However language parsers (either oracle or trained parsers) are not accurate at understanding negation, and as such this approach does not yield correct answers for logically composed questions. The question then arises: can the model answer such questions, if we explicitly train it with data that also contains logically composed questions? For this investigation, we construct two datasets, VQA-Compose and VQA-Supplement, by utilizing annotations from the

VQA dataset, as well as object and caption annotations from COCO [25]. We use these datasets to train the state-of-the-art model LXMERT [43] and perform multiple experiments to test for robustness towards logically composed questions.

After this investigation, we develop our LOL model architecture that jointly learns to answer questions while understanding the type of question and which logical connective exists in the question, through our attention modules, as shown in Figure 3. We further train our model with a novel Fréchet-Compatibility loss that ensures compatibility between the answers to the component questions and the answer of the logically composed question. One key finding is that our models are better than existing models trained on logical questions, with a small deviation from state-of-the-art on VQA test set. Our models also exhibit better *Compositional Generalization* i.e. models trained to answer questions with a single logical connective are able to answer those with multiple connectives.

Our contributions are summarized below:

1. We conduct a detailed analysis of the performance of the state-of-the-art VQA model with respect to logically composed questions,
2. We curate two large scale datasets `VQA-Compose` and `VQA-Supplement` that contain logically composed binary questions.
3. We propose *LOL* – our end-to-end model with dedicated attention modules that answer questions by understanding the logical connectives in questions.
4. We show a capability of answering logically composed questions, while retaining VQA performance.

2   Related Work

Logic in Human Expression: Is logical thinking a natural feature of human thought and expression? Evidence in psychological studies [10,16,11] suggests that infants are capable of logical reasoning, toddlers understand logical operations in natural language and are able to compositionally compute meanings even in complex sentences containing multiple logical operators. Children are also able to use these meanings to assign truth values to complex experimental tasks. Given this, question-answering systems also need to answer compositional questions, and be robust to the manifestation of logical operators in natural language.

Logic in Natural Language Understanding: The task of understanding compositionality in question-answering (QA) can also be interpreted as understanding logical connectives in text. While question compositionality is largely unstudied, approaches in natural language understanding seek to transform sentences into symbolic formats such as first-order logic (FOL) or relational tables [30,47,24]. While such methods benefit from interpretability, they suffer from practical limitations like intractability, reliance on background knowledge, and failure to process noise and uncertainty. [8,39,41] suggest that better generalization can be achieved by learning embeddings to reason about semantic relations, and to simulate FOL behavior [40]. Recursive neural networks have been shown to learn logical semantics on synthetic English-like sentences by using embeddings [9,32].

(a)

Is the lady holding the baby?

Is the man holding the baby?

Is the baby holding the man?

(b)

Are they in a restaurant?

Are they all girls?

Are they in a restaurant and
are they all boys?

Fig. 2: Some questions in `VQA-Supplement` created with adversarial antonyms.

Detection of negation in text has been studied for information extraction and sentiment analysis [31]. [22] have shown that BERT-based models [13,26] are incapable of differentiating between sentences and their negations. Concurrent to our work, [4] show the efficacy of FOL-guided data augmentation for performance improvements on natural language QA tasks that require reasoning. Since our work deals with both vision and language modalities, it encounters a greater degree of ambiguity, thus calling for robust VQA systems that can deal with logical transformations.

Visual Question Answering (VQA) [3] is a large-scale, human-annotated dataset for open-ended question-answering on images. VQA-v2[17] reduces the language bias in the dataset by collecting complementary images for each question-image pair. This ensures that the number of questions in the VQA dataset with the answer "YES" is equal to those with the answer "NO". This dataset contains 204k images from MS-COCO [25], and 1.1M questions.

Cross-modal pre-trained models [43,27,48] have proved to be highly effective in vision-and-language tasks such as VQA, referring expression comprehension, and image retrieval. While neuro-symbolic approaches [29] have been proposed for VQA tasks which require reasoning on synthetic images, their performance on natural images is lacking. Recent work seeks to incorporate reasoning in VQA, such as visual commonsense reasoning [46,14], spatial reasoning [20,21], and by integrating knowledge for end-to-end reasoning [1].

We take a step back and extensively analyze the pivotal task of VQA with respect to various aspects of generalization. We consider a rigorous investigation of a task, dataset, and models to be equally important as proposing new challenges that are arguably harder. In this paper we analyse existing state-of-the-art VQA models with respect to their robustness to logical transformations of questions.

3   The Lens of Logic

A lens magnifies objects under investigation, by allowing us to zoom and focus on desired contents or processes. Our lens of logical composition of questions, allows us to magnify, identify, and analyze the problems in VQA models.

Consider Figure 2(a), where we transform the first question *"Is the lady holding the baby"* by first replacing *"lady"* with an adversarial antonym *"man"* and observe that the system provides a wrong answer with very high probability.

Table 1: Illustration of question composition in `VQA-Compose`, for the same example as in Figure 1. QF: Question Formula, AF: Answer Formula

QF	Question	AF	Answer
Q_1	Is there beer?	A_1	Yes
Q_2	Is the man wearing shoes?	A_2	No
$\neg Q_1$	Is there no beer?	$\neg A_1$	No
$\neg Q_2$	Is the man not wearing shoes?	$\neg A_2$	Yes
$Q_1 \wedge Q_2$	Is there beer and is the man wearing shoes?	$A_1 \wedge A_2$	No
$Q_1 \vee Q_2$	Is there beer or is the man wearing shoes?	$A_1 \vee A_2$	Yes
$Q_1 \wedge \neg Q_2$	Is there beer and is the man not wearing shoes?	$A_1 \wedge \neg A_2$	Yes
$Q_1 \vee \neg Q_2$	Is there beer or is the man not wearing shoes?	$A_1 \vee \neg A_2$	Yes
$\neg Q_1 \wedge Q_2$	Is there no beer and is the man wearing shoes?	$\neg A_1 \wedge A_2$	No
$\neg Q_1 \vee Q_2$	Is there no beer or is the man wearing shoes?	$\neg A_1 \vee A_2$	No
$\neg Q_1 \wedge \neg Q_2$	Is there no beer and is the man not wearing shoes?	$\neg A_1 \wedge \neg A_2$	No
$\neg Q_1 \vee \neg Q_2$	Is there no beer or is the man not wearing shoes?	$\neg A_1 \vee \neg A_2$	Yes

Swapping *"man"* with *"baby"* results in a wrong answer as well. In 2(b) a conjunction of two questions containing antonyms (*girls* vs *boys*) yields a wrong answer. We identify that the ability to answer composite questions created by negation, conjunction and disjunction of questions is crucial for VQA.

We use "closed questions" as defined in [6] to construct logically composed questions. Under this definition, if a closed question has a negative ("NO") answer then its negation must have an affirmative ("YES") answer. Of the three types of questions in the VQA dataset (yes/no, numeric, other), 'yes-no' questions satisfy this requirement. Although, visual questions in the VQA dataset can have multiple correct answers [5], 20.91% of the questions (around 160k) in the VQA dataset are closed questions, i.e. questions with a single unambiguous yes-or-no answer, unanimously annotated by multiple human workers. This allows us to treat these questions as propositions and create a truth table for answers to compose logical questions as shown in Table 1.

3.1   Composite Questions

Let $\mathcal{D}$ be the VQA dataset. For closed questions Q_1 and Q_2 about image $I \in \mathcal{D}$, we define the composite question Q^* composed using connective $\circ \in \{\vee, \wedge\}$, as:

$$Q^* = \widehat{Q_1} \circ \widehat{Q_2}, \qquad where \ \ \widehat{Q_1} \in \{Q_1, \neg Q_1\}, \ \widehat{Q_2} \in \{Q_2, \neg Q_2\}. \qquad (1)$$

3.2   Dataset Creation Process

Using the above definition we create two new datasets by utilizing multiple questions about the same image (`VQA-Compose`) and external object and caption annotations about the image from COCO to create more questions (`VQA-Supplement`).

The seed questions for creating these datasets are all closed binary questions from VQA-v2 [17]. These datasets serve as test-beds, and enable experiments that analyze performance of models when answering such questions.

VQA-Compose: Consider the first two rows in Table 1. Q_1 and Q_2 are two questions about the image in Figure 1 taken from the VQA dataset. Additional questions are composed from Q_1 and Q_2 by using the formulas in Table 1. Thus for each pair of closed questions in the VQA dataset, we get 10 logically composed questions. Using the same train-val-test split as the VQA-v2 dataset [17], we get *1.25 million samples* for our `VQA-Compose` dataset. The dataset is balanced in terms of the number of questions with affirmative and negative answers.

VQA-Supplement: Images in VQA-v2 follow identical train-val-test splits as their source MS-COCO [25]. Therefore, we use the object annotations from COCO to create additional closed binary questions, such as *"Is there a bottle"* for the example in Figure 1. We also create "adversarial" questions about objects, like *"Is there a wine-glass?"* by using an object that is not present in the image (wine-glass), but is *semantically close* to an object in the image (bottle). We use Glove vectors [33] to find the adversarial object with the closest embedding. Following a similar strategy, we also convert captions provided in COCO to closed binary questions, for example *"Does this seem like a man bending over to look inside the fridge"*. Since we know what objects are present in the image, and the captions describe a "true" scene, we are able to obtain the ground-truth answers for questions created from objects and captions. Similar methods for creation of question-answer pairs have previously been used in [37,28].

Thus for every question, we obtain several questions from objects and captions, and use these to compose additional questions by following a process similar to the one for `VQA-Compose`. For each closed question in the VQA dataset, we get 20 additional logically composed questions by utilizing questions created from objects and captions, yielding a total of *2.55 million samples* as `VQA-Supplement`.

3.3   Analytical Setup

In order to test the robustness of our models to logically composed questions, we devise five key experiments to analyse baseline models and our methods. These experiments help us gain insights into the nuances of the VQA dataset, and allow us to develop strategies for promoting robustness.

Effect of Data Augmentation: In this experiment, we compare the performance of models on `VQA-Compose` and `VQA-Supplement` with or without logically composed training data. This experiment allows us to test our hypotheses about the robustness of any VQA model to logically composed questions. We first use models trained on VQA data to answer questions in our new datasets and record performance. We then explicitly train the same models with our new datasets, and make a comparison of performance with the pre-trained baseline.
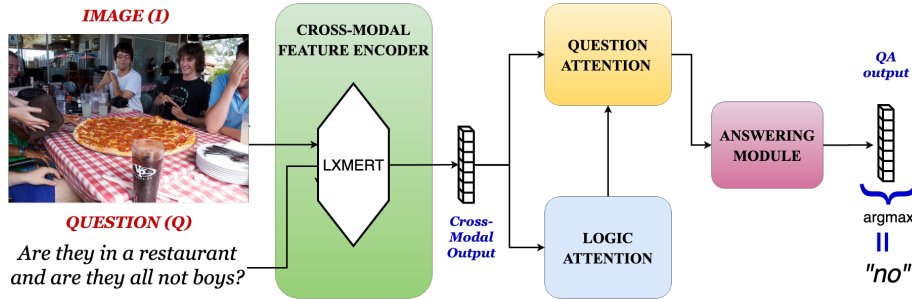
Fig. 3: LOL model architecture showing a cross-modal feature encoder followed by our Question-Attention (q_{ATT}) and Logic Attention (ℓ_{ATT}) modules. The concatenated output of is used by the Answering Module to predict the answer.

Learning Curve: We train our models with an increasing number of logically composed questions and compare performance. This serves as an analysis of the number of logical samples needed by the model to understand logic in questions.

Training only with Closed Questions: In this ablation study, we restrict the training data to only closed questions i.e. "Yes-No" VQA questions, `VQA-Compose` and `VQA-Supplement`, allowing our model to focus solely on closed questions.

Compositional Generalization: We address whether training on closed questions containing single logical operation ($\neg Q_1$, $Q_1 \vee Q_2$) can generalize to multiple operations ($Q_1 \wedge \neg Q_2$, $\neg Q_1 \vee Q_2$). For instance, rows 1 through 6 in Table 1 are *single operation questions*, while rows 7 through 12 are *multi-operation questions*. Our aim is to have models that exhibit such compositional generalization.

Inductive Generalization: We investigate if training on compositions of two questions ($\neg Q_1 \vee Q_2$) can generalize to compositions of more than two questions ($Q_1 \wedge \neg Q_2 \wedge Q_3 \ldots$). This studies whether our models develop an understanding of logical connectives, as opposed to simply learning patterns from large data.

4 Method

In this section. we describe LXMERT [43] (a state-of-the-art VQA model), our Lens of Logic (LOL) model, attention modules which learn the question-type and logical connectives in the question, and the Fréchet-Compatibility (FC) Loss. This section refers to a composition of two questions, but applies to $n \geq 2$ questions.

4.1 Cross-Modal Feature Encoder

LXMERT (Learning Cross-Modality Encoder Representations from Transformers) [43] is one of the first cross-modal pre-trained frameworks for vision-and-language tasks, that combines a strong visual feature extractor [38] with a strong

language model (BERT)[13]. LXMERT is pre-trained for key vision-and-language tasks, on a large corpus of $\sim$9M image-sentence pairs, making it a powerful cross-modal encoder for vision+language tasks such as visual question answering, as compared to other models such as MCAN [45] and UpDn [2], and strong representative baseline for our experiments.

4.2   Our Model: Lens of Logic (LOL)

The design for our LOL model is driven by three key insights:

1. As logically composed questions are closed questions, understanding the type of question will guide the model to answer them correctly.
2. Predicted answers must be compatible with the predicted question type. For instance, a closed question can have an answer that is either "Yes" or "No".
3. The model must learn to identify the logical connectives in a question.

Given these insights, we develop the Question Attention module that encodes the type of question (*Yes-No*, *Number*, or *Other*), and the Logic Attention module that predicts the connectives (*AND, OR, NOT, no connective*) present in the question, and use these to learn representations. The overall model architecture is shown in Figure 3. For every question Q and corresponding image I, we obtain embeddings z_Q and z_I respectively, as well as a cross-modal embedding z_X.

   Question Attention Module (q_{ATT}) takes cross-modal embedding z_x from LXMERT as input, and outputs vector $P_{type} = softmax(\mathbf{q}_{ATT}(z_x))$, representing the probabilities of each question-type. These probabilities are used to get a final representation $\mathbf{z^{type}}$ which combines the features for each question-type.[1]

   Logic Attention Module (ℓ_{ATT}) takes the cross-modal embedding z_X from LXMERT as input, and outputs vector $P^{conn} = \sigma(\ell_{ATT}(z_X))$ which represents the probabilities of each type of connective. We use sigmoid (σ) instead of a softmax, since a question can have multiple connectives. These probabilities are used to combine the features for each type of connective into a final representation $\mathbf{z^{conn}}$ which encodes information about the connectives in the question.

4.3   Loss Functions

We train our models jointly with the loss function given by:

$$\mathcal{L} = (1-\alpha_1-\alpha_2) \cdot \mathcal{L}_{ans} + \alpha_1 \cdot \mathcal{L}_{type} + \alpha_2 \cdot \mathcal{L}_{conn} + \beta \cdot \mathcal{L}_{FC}. \qquad (2)$$

Answering Loss ℓ_{ans} is conditioned on the type of question. We multiply the final prediction vector with the probability and the mask M_i for question-type i. M_i is a binary vector with 1 for every answer-index of type-i and 0 elsewhere:

$$\mathcal{L}_{ans} = \mathcal{L}_{BCE}(\sum_{i=1}^{3} \hat{y} \odot M_i \cdot P_i^{type}, y_{ans}). \qquad (3)$$

Attention Losses: q_{ATT} is trained to minimize a Negative Log Likelihood (NLL) classification loss, ensuring a shrinkage of probabilities of the answer choices of

the wrong type. ℓ_{ATT} is trained to minimize a multi-label classification loss, using Binary Cross-Entropy (BCE) given by:

$$\mathcal{L}_{type} = \mathcal{L}_{NLL}(\text{softmax}(z^{type}), y_{type}), \tag{4}$$

$$\mathcal{L}_{conn} = \mathcal{L}_{BCE}(\sigma(z^{conn}), y_{conn}), \tag{5}$$

where $y_{ans}, y_{type}, y_{conn}$ are labels for answer, question-type and connective.

Fréchet-Compatibility Loss: We introduce a new loss function that ensures compatibility between the answers predicted by the model for the component questions Q_1 and Q_2 and the composed question Q. Let A, A_1, A_2 be the respective answers predicted by the model for Q, Q_1, and Q_2. Q_i can have negation. Then Fréchet inequalities [7,15] provide us with bounds for the probabilities of the answers of the conjunction and disjunction of the two questions:

$$max(0, p(A_1) + p(A_2) - 1) \leq p(A_1 \wedge A_2) \leq min(p(A_1), p(A_2)). \tag{6}$$

$$max(p(A_1), p(A_2)) \leq p(A_1 \vee A_2) \leq min(1, p(A_1) + p(A_2)). \tag{7}$$

We define "Fréchet bounds" b_L and b_R to be the left and right bounds for the triplet A, A_1, A_2, and the "Fréchet Mean" m_A to be the average of the Fréchet bounds; $m_A = (b_L + b_R)/2$. Then, the Fréchet-Compatibility Loss given by:

$$\mathcal{L}_{FC} = (p(A) - \mathbb{1}(m_A > 0.5))^2, \tag{8}$$

ensures that the predicted answer and that determined by m_A match.

4.4   Implementation Details

The LXMERT feature encoder produces a vector z of length 768 which is used by our attention modules, each having sub-networks $\mathbf{f_i}, \mathbf{g_i}$ with 2 feed-forward layers. We first train our models without FC loss. Then we select the best models with a checkpoint of 10 epochs and finetune these further for 3 epochs with FC loss, since the FC loss is designed to work for a model whose predictions are not random. Thus our improvements in accuracy are attributable to the FC Loss and not more training epochs. We utilize the Adam optimizer [23] with a learning rate of $5e$-5, batch size of 32 and train for 20 epochs. Our models are trained on 4 NVIDIA V100 GPUs, and take approximately 24 hours for training 20 epochs.[1]

5   Experiments

We first conduct analytical experiments to test for logical robustness and transfer learning capability. We use three datasets for our experiment: the VQA v2.0 [3] dataset, a combination of VQA and our `VQA-Compose` dataset, and a combination

---

[1] More training details in Supplementary Materials

[2] In all tables, best overall scores are bold, our best scores underlined.

Table 2: Comparison of LXMERT and LOL trained on VQA data, combinations with `Compose`, `Supplement`, and our Frechet-Compatibility (FC) Loss [2]

Model	Trained on	Validation Accuracy (%) ↑			
		VQA	YN	Comp	Supp
LXMERT	VQA	68.94	**86.65**	50.79	50.51
	VQA + Comp	67.85	85.32	85.03	80.85
	VQA + Comp + Supp	68.83	84.83	70.28	85.17
with FC Loss	VQA + Comp + Supp	67.84	84.92	75.31	85.25
LOL (qATT)	VQA	**69.08**	85.32	48.99	50.54
	VQA + Comp	67.51	84.82	84.85	79.62
	VQA + Comp + Supp	68.72	84.99	79.88	87.12
LOL (Full)	VQA + Comp	68.94	85.15	**85.13**	79.02
	VQA + Comp + Supp	68.86	84.87	81.07	87.54
with FC Loss	VQA + Comp + Supp	68.10	84.75	82.39	**87.80**
LXMERT	YN + Comp	-	84.13	84.44	79.39
	YN + Comp + Supp	-	84.09	82.63	88.15
LOL (ℓATT)	YN + Comp	-	85.22	**85.31**	79.87
	YN + Comp + Supp	-	85.26	84.37	**89.00**

Table 3: Validation accuracies (%) for Compositional Generalization and Commutative Property. Note that 50% is random performance.[2]

Model	VQA-Compose		VQA-Supplement		Model	VQA-Compose		VQA-Supplement	
	YN	Single	Multiple	Single	Multiple		$Q_1 \circ Q_2$	$Q_2 \circ Q_1$	$Q_1 \circ Q_2$	$Q_2 \circ Q_1$
LXMERT	85.07	83.95	61.99	86.65	60.00	LXMERT	82.34	80.44	85.57	81.78
LOL	85.12	84.60	66.03	**87.42**	66.05	LOL	84.91	83.64	85.62	83.41

of VQA, `VQA-Compose` and `VQA-Supplement`. The size of the training dataset and the distribution of yes-no, number and other questions is kept the same as the original VQA dataset (∼443k) for fair comparison. Since `VQA-Supplement` uses captions and objects from MS-COCO, we use is to analyze the ability of our models to generalize to a new source of data (MS-COCO) as well as questions containing adversarial objects. After training, our attention modules (q_{ATT} and ℓ_{ATT}) achieve an accuracy of 99.9% on average, showing almost perfect performance when it comes to learning the type of question and the logical connectives present in the question.

5.1   Can't We Just Parse the Question into Components?

Since our questions are a composition of multiple questions, an obvious approach is to split the question into its components, and to discern the logical formula
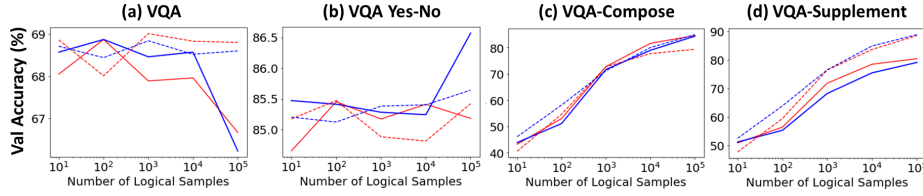
Fig. 4: Learning Curve comparison for models (Red: LXMERT, Blue: LOL) trained on our datasets (solid lines: VQA + `Comp`, dotted lines: VQA + `Comp` + `Supp`)

for composition. The answers to these component questions (predicted by VQA models) can be *re-combined* with the predicted logical formula to obtain the final answer. We use parsers to map components and logical operations to predefined slots in a logical function. The oracle parser uses the ground truth component questions and combines predicted answers using the true formula. However, at test time we do not have access to the true mapping and components. So we train a RoBERTa-Base [26] parser using B-I-O tagging [36] for a Named-Entity Recognition task with constituent questions as entities.[1]

The performance of the oracle parser serves as the upper bound as we have a perfect mapping, with the QA system being the only source of error. The trained parser has an exact-match accuracy of 85%, but only a 72% accuracy in determining the number of operands. The parser has an accuracy of 89% for questions with 3 or less operands, but only 78% for longer compositions. End-to-end (E2E) models do not need to parse questions and hence overcome these hurdles, but do require an understanding of logical operations. Table 4 shows that both oracle and trained parsers when used with LOL outperform parsers with LXMERT, by 6.82%) and 5.60% respectively. The LOL model without using any parsers is better than both LXMERT and LOL with the trained parser by 7.55% and 1.95% respectively.

5.2   Explicit Training with Logically Composed Questions

Can models trained on the VQA-v2 dataset answer logically composed questions? The first section of Table 2 shows that LXMERT, when trained only on questions from VQA-v2 has near random accuracy (∼50%) on our logically composed datasets, thus exhibiting little robustness to such questions.

Can baseline model improve if trained explicitly with logically composed questions questions? We train the models with data containing a combination of samples from VQA-v2, `VQA-Compose`, and `VQA-Supplement`. The accuracy on `VQA-Compose` and `VQA-Supplement` improves, but there is a drop in performance on yes-no questions from VQA. Our models with our attention modules (q_{ATT} and ℓ_{ATT}) are able to retain performance on VQA-v2 while achieving improvements on all validation datasets.
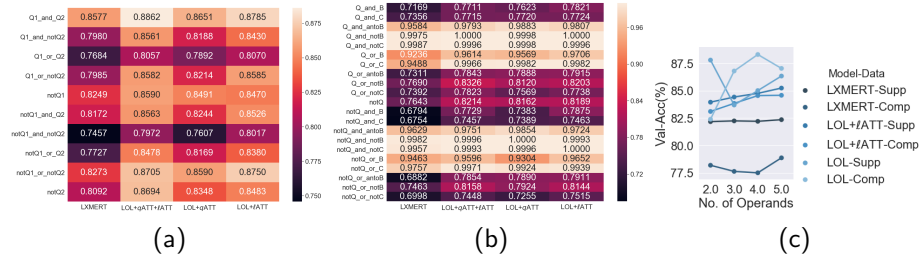
Fig. 5: Accuracy for each type of question in (a) VQA-Compose, (b) VQA-Supplement and for questions with number of operands greater than 2.

5.3   Analysis

Training with Closed Questions only: We analyse the performance of models when trained only with closed questions from VQA, VQA + Comp and VQA + Comp + Supp and see that our model achieves the best accuracy on logically composed questions, as shown in sections 3 and 4 in Table 2. Since we train only closed questions, we do not use our question attention module for this experiment.

Effect of Logically Composed Questions: We increase the number of logical samples in the training data on a log scale from 10 to 100k. As can be seen from the learning curves in Figure 4(a), models trained on VQA + Comp + Supp are able to retain performance on VQA validation data, while those trained only on VQA + Comp data deteriorate. Figure 4(b) shows that our models improve on VQA Yes-No performance after being trained on more logically composed samples, exhibiting transfer learning capabilities. In (c) both our models are comparable to the baseline, but our model shows improvements over the baseline when trained on VQA + Comp + Supp. In (d) for all levels of additional logical questions, our model trained on VQA + Comp + Supp is the best performing. From (c) and (d), we observe that a large number of logical questions are needed during training for the models to learn to answer them during inference. We also see that our model yields the best performance on VQA-Supplement.

Compositional Generalization: To test for compositional generalization, we train models on questions with a maximum of one connective (single) and test on those with multiple connectives. It can be seen from Table 3 that our models are better equipped than the baseline to generalize to multiple connectives and also to be able to generalize from VQA-Compose to Supplement.

Inductive Generalization: We test our models on questions composed with more than two components. Parser-based models have this property by default. As shown by Figure 5c our E2E models outperform the baseline LXMERT.

Table 4: Performance on 'test-standard' set of VQA-v2 and validation set of our datasets. LOL performance is close to SOTA on VQA-v2, but significantly better at logical robustness. *MCAN uses a fixed vocabulary that prohibits evaluation on `VQA-Supplement` which has questions created from COCO captions. #Test-dev scores, since MCAN does not report test-std single-model scores[2]

Model	Parser	Training Data	Test-Std. Accuracy (%) ↑				Val. Accuracy (%) ↑		
			Yes-No	Number	Other	Overall	Compose	Supplement	Overall
MCAN	None	VQA [45]	86.82#	53.26#	60.72#	70.90	52.42	*	*
LXMERT	None	VQA [43]	**88.20**	**54.20**	**63.10**	**72.50**	50.79	50.51	50.65
LOL (qATT)	None	VQA	<u>87.33</u>	<u>54.03</u>	<u>62.40</u>	<u>72.03</u>	48.99	50.54	49.77
LXMERT	Oracle	VQA	88.20	54.20	63.10	72.50	86.38	74.29	80.33
LXMERT	Trained	VQA	88.20	54.20	63.10	72.50	86.35	68.75	77.55
LOL (full)	Oracle	VQA+Ours	86.55	53.42	61.58	71.04	85.79	88.51	87.15
LOL (full)	Trained	VQA+Ours	86.55	53.42	61.58	71.04	82.13	84.17	83.15
LXMERT	None	VQA+Ours	85.23	51.25	60.58	69.78	75.31	85.25	80.28
LOL (qATT)	None	VQA+Ours	86.79	52.66	61.85	71.19	79.88	87.12	83.50
LOL (full)	None	VQA+Ours	86.55	53.42	61.58	71.04	**82.39**	**87.80**	85.10

Commutative Property: Our models have identical answers when the question is composed either as $Q_1 \circ Q_2$ or $Q_2 \circ Q_1$, for logical operation $\circ$, as shown in Table 3. The parser-based models are agnostic to the order of components if the parsing is accurate, while our E2E models are robust to the order.

Accuracy per Category of Question Composition: In Figure 5 we show a plot of accuracy versus question type for each model. Q, Q_1, Q_2 are questions from VQA, B, C are object-based and caption-based questions from COCO respectively. From the results, we interpret that questions such as $Q \wedge antonym(B), Q \wedge \neg B, Q \wedge \neg C$ are easy because the model is able to understand absence of objects, therefore can always answer these questions with a "NO". Similarly, $Q \vee B, Q \vee C$ are easily answered since presence of the object makes the answer always "YES". By simply understanding object presence many such questions can be answered. Figure 5 shows the model has the same accuracy for logically equivalent operations.

5.4   Evaluation on VQA v2.0 Test Data

Table 4 shows the performance the VQA Test-Standard datset. Our models maintain overall performance on the VQA test dataset, and at the same time substantially improve from random performance ($\sim 50\%$) on logically composed questions to 82.39% on `VQA-Compose` and 87.80% on `VQA-Supplement`. This shows that logical connectives in questions can be learned while not degrading the overall performance on the original VQA test set (our models are within $\sim 1.5\%$ of the state-of-the-art on all three types of questions on the VQA test-set).

6   Discussion

Consider the example, *"Is every boy who is holding an apple or a banana, not wearing a hat?"*, humans are able to answer it to be true if and only if each boy who is holding *at least one* of an apple or a banana is not wearing a hat [11]. Natural language contains such complex logical compositions, not to mention ambiguities and the influence of context. In this paper, we focus on the simplest – negation, conjunction, and disjunction. We have shown that existing VQA models are not robust to questions composed with these logical connectives, even when we train parsers to split the question into its components. When humans are faced with such questions, they may refrain from giving binary (Yes/No) answers. For instance, logically, the question *"Did you eat the pizza and did you like it?"* has a negative answer if either of the two component questions has a negative answer. However, humans might answer the same question with the answer *"Yes, but I did not like it"*. While human question-answering is indeed elaborate, explanatory, and clarifying, that is the scope of our future work; here we focus only on predicting a single binary answer.

We have shown how connectives in a question can be identified by enhancing LXMERT encoders with dedicated attention modules and loss functions. We would like to stress on the fact that we do not use knowledge of the connectives during inference, but instead train the network to be aware of it based on cross-modal features, instead of predicting purely based on language model embeddings which fail to capture these nuances. Our work is an attempt to modularize the understanding of logical components to train the model to utilize the outputs of the attention modules. We believe this work has potential implications on logic-guided data augmentation, logically robust question answering, and for conversational agents (with or without images). Similar strategies and learning mechanisms may be used in the future to operate "logically" in the image-space at the level of object classes, attributes, or semantic segments.

7   Conclusion

In this work, we investigate VQA in terms of logical robustness. The key hypothesis is that the ability to answer questions about an image, must be extendable to a logical composition of two such questions. We show that state-of-the-art models trained on VQA dataset lack this. Our solution involves the "Lens of Logic" model architecture that learns to answer questions with negation, conjunction, and disjunction. We provide `VQA-Compose` and `VQA-Supplement`, two datasets containing logically composed questions to serve as benchmarks. Our models show improvements in terms of answering these questions, while at the same time retaining performance on the original VQA test-set.

Acknowledgments

References

1. Aditya, S., Yang, Y., Baral, C.: Integrating knowledge and reasoning in image understanding. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence. pp. 6252–6259. IJCAI'19, AAAI Press (2019), `http://dl.acm.org/citation.cfm?id=3367722.3367926` 4
2. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: CVPR (2018) 8
3. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015) 2, 4, 9
4. Asai, A., Hajishirzi, H.: Logic-guided data augmentation and regularization for consistent question answering. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 5642–5650. Association for Computational Linguistics, Online (Jul 2020), `https://www.aclweb.org/anthology/2020.acl-main.499` 4
5. Bhattacharya, N., Li, Q., Gurari, D.: Why does a visual question have different answers? In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4271–4280 (2019) 5
6. Bobrow, D.G.: Natural language input for a computer problem solving system (1964) 5
7. Boole, G.: An investigation of the laws of thought: on which are founded the mathematical theories of logic and probabilities. Dover Publications (1854) 2, 9
8. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: Advances in neural information processing systems. pp. 2787–2795 (2013) 3
9. Bowman, S.R., Potts, C., Manning, C.D.: Recursive neural networks can learn logical semantics. arXiv preprint arXiv:1406.1827 (2014) 3
10. Carey, S.: Conceptual change in childhood. MIT press (1985) 3
11. Cesana-Arlotti, N., Martín, A., Téglás, E., Vorobyova, L., Cetnarski, R., Bonatti, L.L.: Precursors of logical reasoning in preverbal human infants. Science **359**(6381), 1263–1266 (2018). `https://doi.org/10.1126/science.aao3539`, `https://science.sciencemag.org/content/359/6381/1263` 2, 3, 14
12. Corcoran, J.: Completeness of an ancient logic. The journal of symbolic logic **37**(4), 696–702 (1972) 2
13. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018) 4, 8
14. Fang, Z., Gokhale, T., Banerjee, P., Baral, C., Yang, Y.: Video2commonsense: Generating commonsense descriptions to enrich video captioning. arXiv preprint arXiv:2003.05162 (2020) 4
15. Fréchet, M.: Généralisation du théoreme des probabilités totales. Fundamenta mathematicae **1**(25), 379–387 (1935) 9
16. Gopnik, A., Meltzoff, A.N., Kuhl, P.K.: The scientist in the crib: Minds, brains, and how children learn. William Morrow & Co (1999) 3
17. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6904–6913 (2017) 4, 6

18. Hegel, G.W.F.: Hegel's science of logic (1929) 2
19. Horn, L.R., Kato, Y.: Negation and polarity: Syntactic and semantic perspectives. OUP Oxford (2000) 2
20. Hudson, D.A., Manning, C.D.: Gqa: a new dataset for compositional question answering over real-world images. arXiv preprint arXiv:1902.09506 (2019) 4
21. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2901–2910 (2017) 4
22. Kassner, N., Schütze, H.: Negated lama: Birds cannot fly. arXiv preprint arXiv:1911.03343 (2019) 4
23. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 9
24. Lewis, M., Steedman, M.: Combined distributional and logical semantics. Transactions of the Association for Computational Linguistics **1**, 179–192 (2013). https://doi.org/10.1162/tacl_a_00219, https://www.aclweb.org/anthology/Q13-1015 3
25. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014) 3, 4, 6
26. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019) 4, 11
27. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Advances in Neural Information Processing Systems. pp. 13–23 (2019) 4
28. Malinowski, M., Fritz, M.: A multi-world approach to question answering about real-world scenes based on uncertain input. In: Advances in neural information processing systems. pp. 1682–1690 (2014) 6
29. Mao, J., Gan, C., Kohli, P., Tenenbaum, J.B., Wu, J.: The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision. In: International Conference on Learning Representations (2019), https://openreview.net/forum?id=rJgMlhRctm 4
30. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2. pp. 1003–1011. Association for Computational Linguistics (2009) 3
31. Morante, R., Sporleder, C.: Modality and negation: An introduction to the special issue. Computational linguistics **38**(2), 223–260 (2012) 4
32. Neelakantan, A., Roth, B., McCallum, A.: Compositional vector space models for knowledge base completion. arXiv preprint arXiv:1504.06662 (2015) 3
33. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014) 6
34. Piattelli-Palmarini, M.: Language and learning: the debate between jean piaget and noam chomsky (1980) 2
35. Raju, P.: The principle of four-cornered negation in indian philosophy. The Review of Metaphysics pp. 694–713 (1954) 2

36. Ramshaw, L., Marcus, M.: Text chunking using transformation-based learning. In: Third Workshop on Very Large Corpora (1995), https://www.aclweb.org/anthology/W95-0107 11

37. Ren, M., Kiros, R., Zemel, R.: Exploring models and data for image question answering. In: Advances in neural information processing systems. pp. 2953–2961 (2015) 6

38. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp. 91–99 (2015) 7

39. Riedel, S., Yao, L., McCallum, A., Marlin, B.M.: Relation extraction with matrix factorization and universal schemas. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 74–84. Association for Computational Linguistics, Atlanta, Georgia (Jun 2013), https://www.aclweb.org/anthology/N13-1008 3

40. Rocktäschel, T., Bošnjak, M., Singh, S., Riedel, S.: Low-dimensional embeddings of logic. In: Proceedings of the ACL 2014 Workshop on Semantic Parsing. pp. 45–49 (2014) 3

41. Socher, R., Chen, D., Manning, C.D., Ng, A.: Reasoning with neural tensor networks for knowledge base completion. In: Advances in neural information processing systems. pp. 926–934 (2013) 3

42. Spinoza, B.D.: Ethics, translated by andrew boyle, introduction by ts gregory (1934) 2

43. Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490 (2019) 2, 3, 4, 7, 13

44. Wittgenstein, L.: Tractatus logico-philosophicus. Routledge (1921) 2

45. Yu, Z., Yu, J., Cui, Y., Tao, D., Tian, Q.: Deep modular co-attention networks for visual question answering. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019) 8, 13

46. Zellers, R., Bisk, Y., Farhadi, A., Choi, Y.: From recognition to cognition: Visual commonsense reasoning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6720–6731 (2019) 4

47. Zettlemoyer, L.S., Collins, M.: Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. arXiv preprint arXiv:1207.1420 (2012) 3

48. Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J.J., Gao, J.: Unified vision-language pre-training for image captioning and vqa. In: AAAI. pp. 13041–13049 (2020) 4