Supplementary Material for 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

Abstract. In our paper, we investigated visual question answering (VQA) through the lens of logical transformation. We showed that state-of-the-art VQA models are unable to reliably predict answers for questions composed with logical operations, i.e. negation, conjunction, and disjunction. We introduced new datasets VQA-Compose and VQA-Supplement, created with logical composition and a novel methodology to train models to learn logical operators in questions In this supplementary material, we elaborate upon the following topics:

- Data creation process,
- Dataset analysis,
- Training datasets used for each experiment,
- Additional details about model training and hyper-parameters,
- Additional details about parser models, and
- Further analysis and insights about our results.

1 Dataset Creation

The key idea behind our dataset creation process is to leverage existing annotations from the VQA-v2 dataset [1] and from MS-COCO [3] which is the source of images in VQA-v2. We use questions from VQA-v2, and object annotations and captions from MS-COCO for each image. In order to create logically composed questions, we first filter out the "yes-no" questions which constitute 38% of the VQA dataset. We further filter these by retaining only those yes-no questions with a single valid answer. These questions which are 20% of the VQA data, have an unambiguous answer, chosen unanimously by all human annotators who created the VQA dataset. This satisfies the definition of *"closed questions"* [2] that we use, and are thus the atoms of our data creation process.

We use two closed questions corresponding to the same image to create logically composed questions using the Boolean operators: negation ($\neg$), conjunction ($\wedge$), and disjunction ($\vee$). Since they have a clear unambiguous answer that is either "yes" or "no", we can treat them as Boolean variables, and obtain answers for every new question composed. For negating a question, we follow a template-based procedure negates the question by adding a "no" or "not" before a verb,

* Equal Contribution

Table 1. Examples of question negation. Q denotes the original question from the VQA dataset, $\neg Q$ denotes its negation.

Q	$\neg Q$
Is this an area near the city ?	Is an this area *not* near the city?
Are all the men wearing ties ?	Are all the men *not* wearing ties?
Is there a chair ?	Is there *no* chair?
Do you think it's gonna rain?	Do you think it's *not* gonna rain?

Table 2. Examples of adversarial antonyms for objects. The antonym is chosen such that it is not in the image, but is semantically close to an object in the image

Object	**Adversarial Antonym**
bottle	wine glass
cup	bowl
spoon	fork
surfboard	skateboard
motorcycle	bicycle
sink	toilet

preposition or noun phrase, as shown in Table 1. Note that our data creation method chooses to put a "'not" or "no" either before a preposition, verb, or noun phrase. For instance, *Is this an area near the city?* is transformed to either *Is this not an area near the city?* or *Is this an area not near the city?* randomly. Conjunction and disjunction are straightforward, we add the words "and" and "or" between two closed questions.

1.1 VQA-Compose

`VQA-Compose` is our dataset that is created solely from closed questions in the VQA dataset, by using negation, conjunction and disjunction to compose questions. As shown in Figure 2, we obtain 10 questions for each closed question in the VQA dataset, resulting in a total of 1.25M question-answer-image triplets as our `VQA-Compose` dataset.

1.2 VQA-Supplement

Figure 1 shows examples of captions available in the MS-COCO dataset for images in the VQA-v2 dataset. As shown in Figure 3, we use object annotations and captions from MS-COCO to create questions B and C respectively, using template-based methods. We create `VQA-Supplement` by using logical operators (negation, conjunction, and disjunction) to combine B or C with original questions from VQA-v2.

In addition, we generate questions about adversarial object antonyms. An *adversarial object antonym* is defined as an object that is not present in the

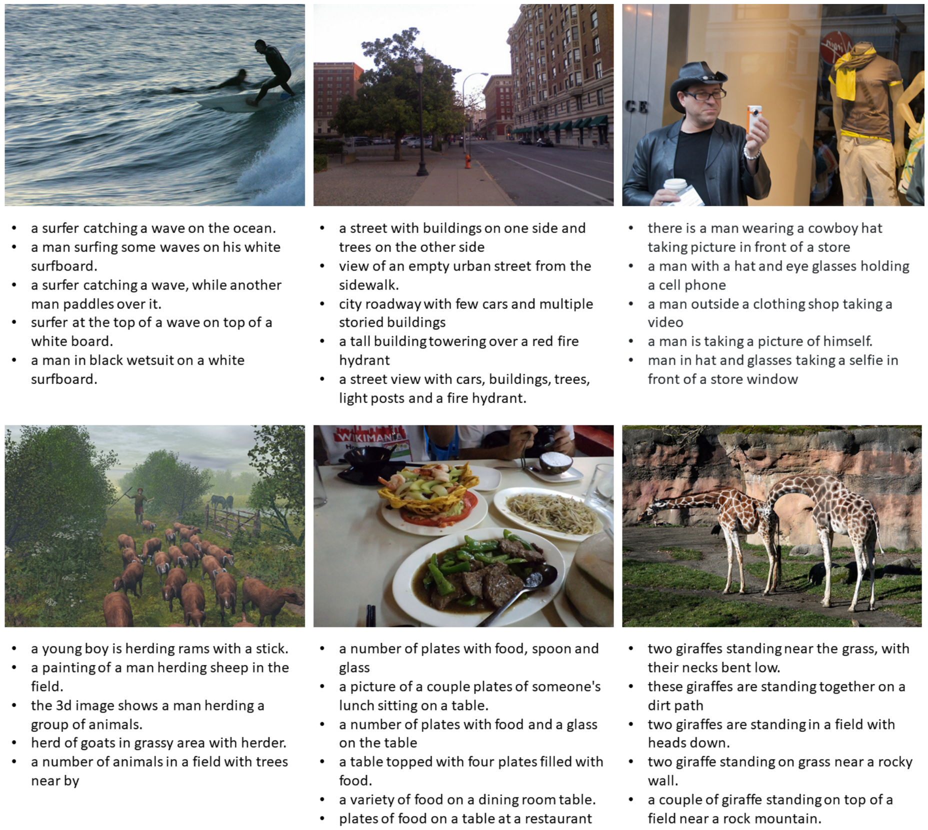


Fig. 1. Examples of captions from COCO for images in the VQA dataset. We convert these captions into questions and use them for our VQA-Supplement dataset

image, but is closest semantically to an object in the image. Examples are shown in Table 2. We use Glove vectors [6] to obtain embeddings of all object class names in the COCO dataset. Then for each image, we find adversarial antonyms using these vectors by using ℓ_2 distance as a metric to sort and select adversarial antonyms. Since the list of objects present in the image is available to us via MS-COCO, we are able to determine the ground-truth answers for object-based questions.

For each question Q we obtain 20 new object-based and caption-based questions. In total, our `VQA-Supplement` dataset contains 2.55M question-answer-image triplets.

2 Dataset Analysis

In this section, we analyze the VQA dataset as well as our new datasets that contain logically composed questions.

IMAGES
from VQA
Validation Set

Questions created in VQA-Compose

QF	AF	Q	A	Q	A	Q	A
Q_1	A_1	Is there a bird in this picture?	No	Do you see animals in picture?	No	Is the man wearing glasses?	Yes
Q_2	A_2	Is the person in the foreground drowning?	No	Is this a busy road?	Yes	Is he wearing a hat?	Yes
$\neg Q_1$	$\neg A_1$	Is there no bird in this picture ?	Yes	Do you not see animals in picture ?	Yes	Is the man not wearing glasses ?	No
$\neg Q_2$	$\neg A_2$	Is the person in the foreground not drowning ?	Yes	Is this a not busy road?	No	Is he not wearing a hat ?	No
$Q_1 \wedge Q_2$	$A_1 \wedge A_2$	Is there a bird in this picture and Is the person in the foreground drowning?	No	Do you see animals in picture and Is this a busy road?	No	Is the man wearing glasses and Is he wearing a hat?	Yes
$Q_1 \vee Q_2$	$A_1 \vee A_2$	Is there a bird in this picture or Is the person in the foreground drowning?	No	Do you see animals in picture or Is this a busy road?	Yes	Is the man wearing glasses or Is he wearing a hat?	Yes
$\neg Q_1 \wedge Q_2$	$\neg A_1 \wedge A_2$	Is there a bird in this picture and Is the person in the foreground not drowning ?	No	Do you not see animals in picture and Is this a busy road?	No	Is the man not wearing glasses and Is he wearing a hat?	No
$\neg Q_1 \vee Q_2$	$\neg A_1 \vee A_2$	Is there a bird in this picture or Is the person in the foreground not drowning ?	Yes	Do you not see animals in picture or Is this a busy road?	No	Is the man not wearing glasses or Is he wearing a hat?	Yes
$Q_1 \wedge \neg Q_2$	$A_1 \wedge \neg A_2$	Is there a bird not in this picture and Is the person in the foreground drowning?	No	Do you see animals in picture and Is this a not busy road ?	No	Is the man wearing glasses and Is he not wearing a hat ?	No
$Q_1 \vee \neg Q_2$	$A_1 \vee \neg A_2$	Is there a bird not in this picture or Is the person in the foreground drowning?	Yes	Do you see animals in picture or Is this a not busy road ?	No	Is the man wearing glasses or Is he not wearing a hat ?	Yes
$\neg Q_1 \wedge \neg Q_2$	$\neg A_1 \wedge \neg A_2$	Is there a bird not in this picture and Is the person in the foreground not drowning ?	Yes	Do you not see animals in picture and Is this a not busy road ?	No	Is the man not wearing glasses and Is he not wearing a hat ?	No
$\neg Q_1 \vee \neg Q_2$	$\neg A_1 \vee \neg A_2$	Is there a bird not in this picture or Is the person in the foreground not drowning ?	Yes	Do you not see animals in picture or Is this a not busy road ?	Yes	Is the man not wearing glasses or Is he not wearing a hat ?	No

Fig. 2. Some examples from our `VQA-Compose dataset`. We show all 10 types of new questions created by original questions Q_1 and Q_2 and the corresponding answers. Q, A, QF, AF denote question, answer, question-formula, and answer-formula respectively. anto(B) represents the adversarial antonym of objects in present in the image.

2.1 Question Length

The average length of questions in VQA-v2 [1] is **6.1 words**. Our datasets have a average length of **12.25 words** for `VQA-Compose` and **15.17** for `VQA-Supplement`. This is longer than VQA-v2 since each of our logically composed questions is made up of multiple component questions.

2.2 Types of Answers

The VQA dataset contains a fixed vocabulary of answers. We obtained the Glove [6] embeddings of these answers, and performed k-means clustering on these embeddings to obtain 50 clusters. We show examples of some of these clusters in Table 3. It can be observed that similar answers, such as those belonging a common category such as *food* or *sports* appear in the same cluster.

Objects (B)
person, cup, cell phone

Captions (C)
- *a man outside a clothing shop taking a video*
- *a man with a hat and eye glasses holding a cell phone*

Questions created in VQA-Supplement

QF	AF	Q	A
Q	A	Is he wearing a hat?	Yes
$\neg Q$	$\neg A$	Is he not wearing a hat?	No
$Q \wedge B$	A	Is he wearing a hat and is there a cell phone?	Yes
$Q \vee B$	$\top$	Is he wearing a hat or is there a cell phone?	Yes
$Q \wedge anto\,(B)$	$\bot$	Is he wearing a hat and is there a bowl?	No
$Q \vee anto(B)$	A	Is he wearing a hat or is there a bowl?	Yes
$Q \wedge C$	A	Is he wearing a hat and is this a man outside a clothing shop taking a video?	Yes
$Q \vee C$	$\top$	Is he wearing a hat or is this a man outside a clothing shop taking a video?	Yes
$Q \wedge \neg B$	$\bot$	Is he wearing a hat and is there no cell phone?	No
$Q \vee \neg B$	A	Is he wearing a hat or is there no cell phone?	Yes
$\neg Q \wedge B$	$\neg A$	Is he not wearing a hat and is there a cell phone?	No
$\neg Q \vee B$	$\top$	Is he not wearing a hat or is there a cell phone?	Yes
$\neg Q \wedge \neg B$	$\bot$	Is he not wearing a hat and is there no cell phone?	No
$\neg Q \vee \neg B$	$\neg A$	Is he not wearing a hat or is there no cell phone?	No
$\neg Q \wedge anto(B)$	$\bot$	Is he not wearing a hat and is there a bowl?	No
$\neg Q \vee anto(B)$	$\neg A$	Is he not wearing a hat or is there a bowl?	No
$Q \wedge \neg C$	$\bot$	Is he wearing a hat and is it not a man with a hat and eye glasses holding a cell phone?	No
$Q \vee \neg C$	A	Is he wearing a hat or is it not a man with a hat and eye glasses holding a cell phone?	Yes
$\neg Q \wedge C$	$\neg A$	Is he not wearing a hat and is this a man outside a clothing shop taking a video?	No
$\neg Q \vee C$	$\top$	Is he not wearing a hat or is this a man outside a clothing shop taking a video?	Yes
$\neg Q \wedge \neg C$	$\bot$	Is he not wearing a hat and is it not a man with a hat and eye glasses holding a cell phone?	No
$\neg Q \vee \neg C$	$\neg A$	Is he not wearing a hat or is it not a man with a hat and eye glasses holding a cell phone?	No

Fig. 3. Some examples from our `VQA-Supplement dataset`. We show all 20 types of new questions created by original questions Q_1 and Q_2 and the corresponding answers. Q, A, QF, AF denote question, answer, question-formula, and answer-formula respectively. $\top, \bot$ are the standard Boolean symbols for top and bottom (true and false)

This shows that Glove embeddings of these answers preserve a notion of similarity. Note that the cluster names in Table 3 are assigned by humans after clustering is complete, for the sake of clarity and illustration, and does not play a role in the clustering process. It is interesting to know that our cluster categories are similar to "knowledge categories" obtained in OK-VQA [5]. The categories in OK-VQA are annotated by human workers in Amazon Mechanical Turk.

Table 3. Selected results of k-means clustering on the Glove embeddings of answers in VQA. k=50.

Cluster Name	Cluster Members
Food	'cooking', 'fast food', 'dishes', 'serving', 'grill', 'pizza hut', 'pizza box', 'lunch', 'restaurant', 'cafe', 'dinner', 'dairy', 'deli', 'menu', 'breakfast', 'cat food', 'burrito', 'food', 'dog food', 'eaten', 'burger', 'french fries', 'food processor', 'pizza cutter', 'grocery store', 'chef', 'pizza', 'vegetarian', 'eat', 'cook', 'food truck', 'chips', 'burgers', 'grocery', 'on pizza', 'eating', 'bar', 'sushi', 'sandwich', 'sandwiches', 'bars'
Geography, Language, Ethnicity	'china', 'thailand', 'america', 'american', 'africa', 'mexican', 'indians', 'russian', 'arabic', 'caucasian', 'american flag', 'german', 'russia', 'oriental', 'japan', 'hispanic', 'british', 'american airlines', 'asian', 'african american', 'italian', 'virgin', 'chinese', 'spanish', 'india', 'thai', 'japanese', 'asia', 'brazil', 'french', 'african', 'persian', 'english'
Flowers, Plants	'tulip', 'weeds', 'windowsill', 'tree branch', 'daffodils', 'carnations', 'elm', 'fern', 'grass', 'roses', 'garden', 'wreath', 'trees', 'pine', 'carnation', 'evergreen', 'sunflowers', 'tree', 'palm tree', 'ivy', 'palm', 'lily', 'iris', 'willow', 'christmas tree', 'vase','bamboo', 'tulips', 'rose', 'bushes', 'lilac', 'dandelions', 'plant', 'orchid', 'flowers', 'lilies', 'vines', 'daisy', 'cactus', 'palm trees', 'flower', 'floral', 'branches', 'bark', 'maple leaf', 'leaf', 'daffodil'
Fruits	'mango', 'apples', 'juice', 'cherries', 'strawberries', 'ginger', 'watermelon', 'cane', 'cherry', 'sweet', 'peach', 'organic', 'cantaloupe', 'orange juice', 'banana split', 'ripe', 'lemonade', 'grape', 'fruit', 'sunflower', 'smoothie', 'coconut', 'strawberry', 'banana peel', 'peaches', 'sesame seeds', 'fresh', . . . , 'mint', 'lemons', 'pineapple', 'oranges', 'grapes', 'salt and pepper', 'grapefruit', 'almonds', 'blueberry', 'kiwi'
Birds	'crows', 'pelicans', 'seagull', 'squirrel', 'finch', 'feathers', 'sparrow', 'stork', 'duck', 'parrots', 'rooster', 'eagle', 'bird feeder', 'peacock', 'bird', 'birds', 'goose', 'pigeon', 'crow', 'pigeons', 'owl', 'hummingbird', 'feeder', 'hawk', 'cranes', 'geese', 'flamingo', 'cardinal', 'nest', 'swan', 'ducks', 'parakeet', 'seagulls', 'parrot', 'woodpecker', 'swans', 'pelican'
Sports	'tennis shoes', 'playing game', 'playing baseball', 'tennis', 'baseball bat', 'tennis court', 'football', 'soccer', 'playing video game', 'sports', 'tennis racket', 'baseball uniform', 'team', 'bowling', 'hockey', 'play', 'baseball glove', 'goalie', 'playing tennis', 'badminton', 'playing frisbee', 'tennis player', 'rugby', 'soccer field', 'play tennis', 'soccer ball', 'athletics', 'basketball', . . .
Dog Breeds	'puppy', 'mutt', 'pomeranian', 'dogs', 'dachshund', 'bulldog', 'cocker spaniel', 'schnauzer', 'rottweiler', 'pitbull', 'pug', 'corgi', 'golden retriever', 'german shepherd', 'clydesdale', 'greyhound', 'boxer', 'kitten', 'cat', 'chihuahua', 'dog', 'husky', 'leash', 'terrier', 'dalmatian', 'thoroughbred', 'shepherd', 'sheepdog', 'collie', 'poodle', 'tabby', 'labrador', 'meow', 'beagle', 'calico', 'shih tzu', 'siamese'
Colors	'yellow and red', 'white and blue', 'green and red', 'neon', 'red bull', 'silver and red', 'blue', 'opaque', 'pink and blue', 'orange and yellow', 'black and brown', 'gray and white', 'brown and white', 'blue and black', 'maroon', 'yellow', 'silver', 'gray and red', 'orange and black', 'white and brown', 'black and red', 'black and yellow', 'green', 'purple', 'red and silver', 'colored', 'white and gray', 'black and gray'
Sports Teams	'dodgers', 'mariners', 'mets', 'cardinals', 'braves', 'yankees', 'phillies', 'orioles'
Vegetables	'cauliflower', 'sliced', 'lettuce', 'celery', 'parsley', 'basil', 'squash', 'peppers', 'beets', 'sesame', 'cucumber', 'onion', 'asparagus', 'carrots', 'mushrooms', 'mustard', 'beans', 'broccoli and carrots', 'carrot', 'cilantro', 'cabbage', 'tomato', 'feta', 'veggies', 'avocado', 'peas', 'garlic', 'zucchini', 'pepper', 'vegetables', 'potatoes', 'tomatoes', 'radish',
Bathroom	'toothbrushes', 'lotion', 'washing', 'toiletries', 'faucet', 'mouthwash', 'towel', 'urinal', 'above toilet', 'toothpaste', 'soap', 'pooping', 'bathtub', 'bathing', 'tub', 'drain', 'toilet brush', 'pee', 'shampoo', 'towels', 'on toilet', 'shower', 'bidet', 'toilet paper', 'peeing', 'laundry', 'toilets', 'shower head', . . .
Clothes	'life jacket', 'hat', 'fabric', 'shirts', 'apron', 'bathing suit', 'adidas', 'belt', 'pocket', 'sweater', 't shirt', 'slacks', 'jeans', 'zipper', 'vests', 'bandana', 'costume', 'jackets', 'hoodie', 'strap', 'jacket', 'shoes', 'bow tie', 'pockets', 'yarn', 'denim', 'socks', 't shirt and jeans', 'khaki', 'tuxedo', 'shirt', 'robe', 'swimsuit', 'sleeve', 'overalls', 'uniform', 'cap', 'clothing', 'camouflage', 'fedora', 'suits', 'boots', . . .

Table 4. Training dataset distribution and sizes, for explicit training with new data. Note that training dataset sizes are consistent with the VQA dataset.

Training Datasets	**Proportion of datasets (%)**					**Training Samples**
	VQA-Other	VQA-Number	VQA-YesNo	`Comp`	`Supp`	
VQA	50	12	38	0	0	443754
VQA+Comp	50	12	19	19	0	443754
VQA+Comp+Supp	50	12	12.66	12.66	12.66	443754

Table 5. Training datasets distribution and sizes, for the experiment for understanding the effect of logically composed questions. We progressively add more logical samples, and get the learning curve as shown in the paper.

Training Datasets	**Proportion of samples (%)**					**Training Samples**
	VQA-Other	VQA-Number	VQA-YesNo	`Comp`	`Supp`	
VQA	50	12	38	0	0	443754
VQA + Comp (10)	49.999	11.999	37.999	0.002	0	443764
VQA + Comp (100)	49.989	11.997	37.991	0.022	0	443854
VQA + Comp (1k)	49.888	11.973	37.914	0.225	0	444754
VQA + Comp (10k)	48.898	11.736	37.162	2.204	0	453754
VQA + Comp (100k)	40.805	9.793	31.011	18.391	0	543754
VQA + Comp (10) + Supp (10)	49.998	11.999	37.998	0.002	0.002	443774
VQA + Comp (100) + Supp (100)	49.977	11.995	37.983	0.022	0.022	443954
VQA + Comp (1k)+ Supp (1k)	49.776	11.946	37.829	0.224	0.224	445754
VQA + Comp (10k)+ Supp (10k)	47.844	11.483	36.361	2.156	2.156	463754
VQA + Comp (100k)+ Supp (100k)	34.466	8.272	26.194	15.534	15.534	643754

3 Training Data for Our Experiments

For each experimental setting, we train our models with a dataset containing questions from VQA, `VQA-Compose`, and `VQA-Supplement`. The proportions of these samples in the training data depends upon the specific experiment performed. For each of our experiments we use the same train-validation-test splits as in the VQA-v2 and COCO datasets. In this section, we explain our training datasets in detail for each experiment, analysis, and ablation study.

3.1 Explicit Training with new data

In this experiment, we investigate if existing models trained on VQA data are able to answer questions in `VQA-Compose` and `VQA-Supplement`. We compare this with the LXMERT model [7] trained explicitly with our new data, and also with our models that use the attention modules for question-type and connective-type. For a fair comparison, we restrict the size of training dataset to the original size of the VQA training dataset ($443,754$ samples). We also use the same proportion

Table 6. Training datasets distribution and sizes, for training with logical questions with a maximum of one connective.

Training Datasets	**Proportion of samples (%)**					**Training Samples**
	VQA-Other	VQA-Number	VQA-YesNo	`Comp-Single`	`Supp-Single`	
YesNo	0	0	100	0	0	168626
YesNo + Comp	0	0	50	50	0	337253
YesNo + Comp + Supp	0	0	33.33	33.33	33.33	505879

of question-types as in VQA (38% yes-no, 12% number, and 50% other questions), as shown in Table 4. This allows us to improve the diversity of yes-no questions, by incorporating yes-no questions from `VQA-Compose` and `VQA-Supplement`.

3.2 Training with Closed Questions only

For this experiment, we evaluate the models when trained only on closed questions, under three settings:

1. yes-no questions from VQA
2. yes-no questions from VQA along with an equal number of questions from `VQA-Compose`,
3. yes-no questions from VQA along with an equal number of questions from `VQA-Compose` and `VQA-Supplement`

This allows us to compare the capability of models to answer different types of yes-no questions such as the original questions from VQA, logical compositions in `VQA-Compose`, and logical compositions with object and caption-based questions in `VQA-Supplement`.

3.3 Effect of Logically Composed Questions

In this experiment, we progressively add logically composed questions to the training data, and analyze the learning curve with respect to the number of logical samples We add 10, 100, $1k$, $10k$, and $100k$ samples from `VQA-Compose` or both `VQA-Compose` and `VQA-Supplement`. The training set distribution in shown in Table 5. This allows us to understand how many additional logically composed questions are needed for our models to become robust.

3.4 Compositional Generalization

In this experiment, our aim is to train models on questions that contain a single logical connective (*and, or, not*) or no connective at all (original yes-no questions in VQA), and to test their performance on questions with more than one connective. To do so, we restrict our training data to such single-connective questions as shown in Table 6

Table 7. Hyper-Parameters for training LXMERT and our models

Hyper-Parameters	**Model**
Batch Size	32
Learning Rate	5e-5
Dropout	0.1
Language Layers	9
Cross-Modality Layer	5
Object Relation Layers	5
Optimizer	BertAdam
Warmup	0.1
Max Gradient Norm	5.0
Max Text Length	20

Table 8. Precision-Recall and F1-Scores for the RoBERTa-based NER parser

Operands	**Precision**	**Recall**	**F1-Score**
2	84.98	86.69	85.83
3	81.55	83.62	82.57
4	81.63	83.72	82.66
5	76.29	79.45	77.84

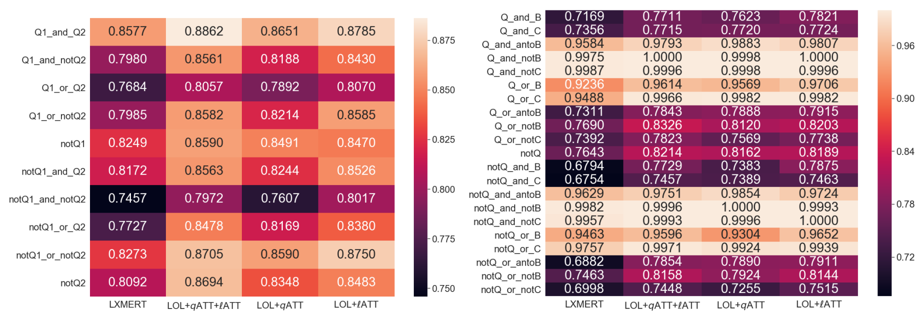
4 Model Architectures and Training Settings

We train our models and baseline LXMERT [7] model with the hyper-parameters in Table 7, chosen from the median of 5 random seeds. The length of cross-modal embeddings produced by LXMERT for each question-image pair is 768. We utilize this as input to our attention modules $\mathbf{q}_{ATT}$ and ℓ_{ATT}. The hidden layers of these attention modules have a size of 2×768. The answering module uses the outputs of these modules to predict softmax answer probabilities.

5 Parser Training and Results

One of our baselines involves using a parser to split a question into its components, answer them separately, and combine the answers logically to get the final answer. We use the RoBERTa-Base language model [4] and train it for the Named-Entity Recognition (NER) task. We modify the RoBERTa-NER model from the Huggingface Transformers [8] framework. We create our parser dataset using the constituent questions as target entities and the original question as the input text. The sequence is classified using B-I-O *(Beginning-Inside-Outside)* [4] tagging scheme, where all constituent tokens are predicted to be tagged as B-Const, I-Const and the connectives are tagged as O. [1] There is only one entity class.

[1] "Const" refers to constituent.



(a) (b)

Fig. 4. Accuracy for each type of question in (a) VQA-Compose, (b) VQA-Supplement and for questions with number of operands greater than 2.

Table 9. Accuracies on each type of question in `VQA-Compose` by each model. QF is Question Formula

QF	LXMERT	LXMERT+ℓ_{ATT}	LXMERT+$\mathbf{q}_{ATT}$	LXMERT+$\mathbf{q}_{ATT}$+ℓ_{ATT}
$\neg Q_1$	85.39	85.55	84.78	86.43
$\neg Q_2$	84.38	85.45	84.94	86.08
$Q_1 \wedge Q_2$	81.50	87.77	87.66	87.77
$Q_1 \vee Q_2$	85.26	81.58	80.54	80.97
$Q_1 \wedge \neg Q_2$	85.71	85.77	84.45	85.02
$Q_1 \vee \neg Q_2$	87.12	86.22	85.98	85.53
$\neg Q_1 \wedge Q_2$	85.10	85.34	84.83	85.53
$\neg Q_1 \vee Q_2$	80.76	78.92	83.79	84.75
$\neg Q_1 \wedge \neg Q_2$	87.98	86.59	79.77	81.32
$\neg Q_1 \vee \neg Q_2$	87.12	85.42	87.42	87.74

We train the model for 20 epochs, with a batch size of 32, and learning rate of 1e-5. The results of our parser are shown in Table 8. It can be observed that the performance of the parser deteriorates as the number of operands in the question increases. This is a major drawback of parser-based methods.

6 Analysis of Results

We provide accuracies of all four models as a heat-map in Figure 4, and also in Tables 9 and 10. We have two key observations.

In Figure 4a, we observe that for all models, the two hardest question categories are $Q_1 \vee Q_2$ and $\neg Q_1 \wedge \neg Q_2$, while the two easiest categories are $Q_1 \wedge Q_2$ and $\neg Q_1 \vee \neg Q_2$. Using DeMorgan's laws to rewrite these logical formulas, we see that

Table 10. Accuracies on each type of question in `VQA-Supplement` by each model

QF	LXMERT	LXMERT+ℓ_{ATT}	LXMERT+q_{ATT}	LXMERT+q_{ATT}+ℓ_{ATT}
Q	82.27	82.3	82.77	82.34
$Q \wedge B$	78.03	77.92	78.16	78.36
$Q \vee B$	95.51	96.79	97.06	96.74
$Q \wedge anto(B)$	95.64	97.55	98.07	96.72
$Q \wedge C$	81.22	82.07	81.67	81.67
$Q \vee C$	99.84	99.89	99.84	99.89
$Q \wedge \neg B$	99.96	99.93	99.98	99.89
$Q \vee \neg B$	82.39	82.54	82.09	81.69
$\neg Q \vee B$	95.08	96.52	96.52	95.51
$\neg Q \wedge \neg B$	99.89	99.84	99.91	99.75
$\neg Q \wedge anto(B)$	94.86	97.91	97.15	97.42
$Q \wedge \neg C$	99.91	99.91	99.98	99.87
$Q \vee \neg C$	82.45	82.21	82.3	81.46
$\neg Q \vee C$	99.80	99.91	99.75	99.82
$\neg Q \wedge \neg C$	99.84	99.87	99.89	99.78
$\neg Q$	80.30	81.62	81.78	80.84
$Q \vee anto(B)$	77.92	77.83	79.13	78.43
$\neg Q \wedge B$	76.27	76.90	78.88	77.31
$\neg Q \vee \neg B$	79.73	81.42	81.49	81.17
$\neg Q \vee anto(B)$	75.62	77.33	79.22	77.92
$\neg Q \wedge C$	78.95	81.26	81.11	80.18
$\neg Q \vee \neg C$	79.87	80.77	81.51	80.61

the two hardest categories are:

$$\mathbf{Q_1} \vee \mathbf{Q_2} \quad , \quad \neg(\mathbf{Q_1} \vee \mathbf{Q_2}),$$

while the two easiest categories are:

$$\mathbf{Q_1} \wedge \mathbf{Q_2} \quad , \quad \neg(\mathbf{Q_1} \wedge \mathbf{Q_2}).$$

Figure 4b provides similar insights. Note that since questions B and C are composed from factually valid statements (about objects in the image, or from valid caption describing a scene), the answers to these questions are always "Yes". Thus answers to any question that uses a disjunction ("or") to combine B, C with another question, is always "Yes". Similarly answers to $\neg B, \neg C, anto(B)$ are always "No". Thus answers to any question that uses a conjunction ("and") to combine $\neg B, \neg C, anto(B)$ with another question, is always "No". These question categories are $Q \vee B, Q \vee C, \neg Q \vee B, \neg Q \vee C$, and $Q \wedge \neg B, Q \wedge \neg C, Q \wedge anto(B), \neg Q \wedge \neg B, \neg Q \wedge \neg C$, and $\neg Q \wedge anto(B)$.

It is interesting to note that questions about adversarial objects are relatively harder to answer for any category and any model, than the questions about objects present in the image. Thus we see that answering questions about objects in the image is much easier than other categories for each model.

Following a similar trend, we observe a difficulty in answering questions which use conjunction ("and") to combine B, C with another question, or which use disjunction ("and") to combine $\neg B, \neg C, anto(B)$ with another question. This is because the answer to these questions changes according to the sample and depends on the answer to the question Q, and cannot be simply "explained away".

References

1. 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) 1, 4
2. Bobrow, D.G.: Natural language input for a computer problem solving system (1964) 1
3. 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) 1
4. 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) 9
5. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3195–3204 (2019) 5
6. 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) 3, 4
7. Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490 (2019) 7, 9
8. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Brew, J.: Huggingface's transformers: State-of-the-art natural language processing. ArXiv **abs/1910.03771** (2019) 9