Revisiting Visual Question Answering Baselines

Allan Jabri, Armand Joulin, Laurens van der Maaten
Facebook AI Research

1. Introduction

- Study of Visual Question Answering has led to the exploration of complex models designed to do multi-modal 'reasoning' and 'attention'.
- We show that a basic approach designed to **exploit answer bias** performs well despite being much simpler:
  - reaches **state-of-the-art** performance on Visual7W Telling.
  - performs competitively on VQA Real multiple choice.
- Surprisingly, the model does 'well' **even when missing** the image or question.

2. Task

Visual Question Answering involves producing the correct text answers given a text question about an image. We consider the **multiple-choice** subtask, in which a smaller set of candidate answers is provided at test time.


What color is the jacket?
-Red and blue.
-Yellow.
-Black.
-Orange.


How many cars are parked?
-Four.
-Three.
-Five.
-Six.

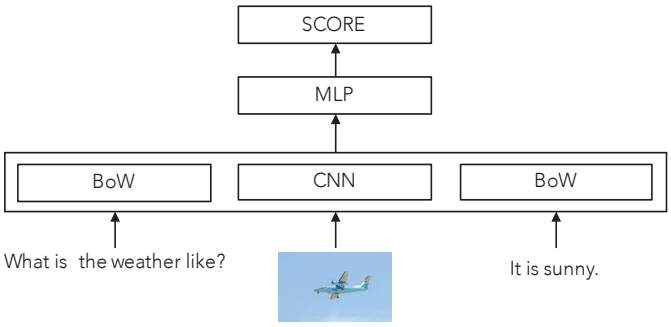

What event is this?
-A wedding.
-Graduation.
-A funeral.
-A picnic.

When is this scene taking place?
-Day time.
-Night time.
-Evening.
-Morning.

We experiment with two datasets:

- Visual7W Telling (*Zhu et al*):
  - $69,817$ train, $28,020$ val, $42,031$ test.
  - Negatives are human-generated.
  - Each question has four answer choices.
  - Performance is measured by accuracy.
- VQA Real multiple-choice (*Agrawal et al*):
  - $248,349$ train, $121,512$ val, $244,302$ test.
  - Negatives are **not** human-generated.
  - Each question has 18 answer choices.
  - $accuracy = \max(\frac{\#human}{3}, 1)$

3. MLP



Our model predicts correctness of an Image-Question-Answer triplet.

- **Loss:** We minimize the binary logistic loss of predicting triplet correctness: $L(x, y) = -y \log f(x) - (1 - y) \log(1 - f(x))$
- **Optimization:** We fit model parameters with SGD, using 25 examples per batch, sampling 2 neg. for each pos., for 300 epochs.
- **Representation:**
  - **Images:** penultimate layer of a ResNet-101 trained on Imagenet.
  - **Text:** BoW (average) of pre-trained word2vec embeddings.
  - All features are **off-the-shelf**, not finetuned, and l_2 normalized.

4. Comparison with the State-of-the-Art

Visual7W Telling

Method	What	Where	When	Who	Why	How	Overall
LSTM (Q, I) [15]	48.9	54.4	71.3	58.1	51.3	50.3	52.1
LSTM-Att [8]	51.5	57.0	75.0	59.5	55.5	49.8	55.6
MCB + Att [21]	60.3	70.4	79.5	69.2	58.2	51.1	62.2
Bilinear (A, Q, I)	60.4	72.3	78.0	71.6	63.0	54.8	63.6
MLP (A)	47.3	58.2	74.3	63.6	57.1	49.6	52.9
MLP (A, Q)	54.9	60.0	76.8	66.0	64.5	54.9	58.5
MLP (A, I)	60.8	74.9	81.9	70.3	64.4	51.2	63.8
MLP (A, Q, I)	**64.5**	**75.9**	**82.1**	**72.9**	**68.0**	**56.4**	**67.1**

VQA Real

Method	Yes/No	Number	Other	All
Two-Layer LSTM [5]	80.6	37.7	53.6	63.1
Region selection [23]	77.2	33.5	56.1	62.4
Question-Image Co-Attention [22]	80.0	**39.5**	59.9	66.1
MCB [21]*	–	–	–	65.4
MCB + Att + GloVe + Genome [21]*	–	–	–	**69.9**
Multi-modal Residual Network [27]	–	–	–	69.3
MLP (A, Q, I)	**80.8**	17.6	**62.0**	65.2

Transfer from VQA to V7W

Model	Method	What	Where	When	Who	Why	How	Overall
MLP (A, Q)	Transfer	44.7	38.9	32.9	49.6	45.0	27.3	41.1
MLP (A, I)	Transfer	28.4	26.6	44.1	37.0	31.7	25.2	29.4
MLP (A, Q, I)	Transfer	58.7	61.7	41.7	60.2	53.2	29.1	53.8
	Finetune	**66.4**	**77.1**	**83.2**	**73.9**	**70.7**	**56.7**	**68.5**

5. Error Analysis

Type	A	AQ	AI	AQI
Counting (11.1)	50.3	55.2	49.8	56.0
Color (13.5)	37.9	52.0	49.3	57.2
Cause (6.3)	57.1	64.5	64.4	68.0
Action (3.0)	59.7	63.5	75.5	77.3
Spatial (4.1)	45.2	51.1	53.1	54.9
Shape (0.5)	45.8	51.1	46.8	54.2
Holding (1.9)	64.4	66.0	69.7	71.7
Time (0.7)	50.0	54.0	62.8	62.8


What is behind the photographer?
-A bus.
-A dump truck.
-A duck.
-A plate of food.


What color leaves are on the tree behind the elephant on the left of the photo?
-Red.
-Orange.
-Green.
-Brown.


What is the man doing?
-Surfing.
-Singing.
-Working.
-Playing.


What is the man doing?
-Golfing.
-Playing tennis.
-Walking.
-Biking.


Why is the ground white?
-Snow.
-Sand.
-Stones.
-Concrete.


Why is his arm up?
-To serve the tennis ball.
-About to hit the ball.
-Reaching for the ball.
-Swinging his racket.


What is the color of the tree leaves?
-Green.
-Brown.
-Orange.
-Red.


What is the color of the train?
-Green.
-Yellow.
-Black.
-Red.

What shape is this sign?
-Octagon.
-Oval.
-Hexagon.
-Square.


What shape is the clock?
-Cube.
-Circle.
-Oval.
-Rectangle.

6. Exploiting Answer Similarity

The model can exploit answer similarity. NN by BoW cosine similarity:

During the daytime.		On a tree branch.		Double decker bus.	
During daytime.	0.97	On the tree branch.	0.98	Doubledecker bus.	0.99
Outside, during the daytime.	0.92	The tree branch.	0.95	A double decker bus.	0.99
Inside, during the daytime.	0.92	Tree branch.	0.95	A doubledecker bus.	0.99
In the daytime.	0.91	A tree branch.	0.95	Double Decker bus.	0.99
In the Daytime.	0.91	Tree branches.	0.87	A red double decker bus.	0.88
During the daytime hours.	0.91	On a tree.	0.86	A double level bus.	0.87
Durin the daytime.	0.89	On tree.	0.86	Double decker tour bus.	0.85
The daytime.	0.89	Eating a tree branch.	0.85	It is a double decker bus.	0.84
In daytime.	0.89	On the tree.	0.83	A Bus.	0.77
During daytime hours.	0.88	On a branch.	0.81	Volkswagon bus.	0.77
During the nighttime.	0.83	The tree.	0.81	A bus.	0.77
At daytime.	0.83	In a tree.	0.81	Bus and.	0.77
Daytime.	0.81	A Tree.	0.80	Doubledecker bus.	0.77
During nighttime.	0.79	Cottonwood tree.	0.80	Bus ischia.	0.77
It occurs in the daytime.	0.77	Tree.	0.80	A countryline bus..	0.77
It is in the daytime.	0.77	A ficus tree.	0.80	Bus.	0.77

Comparing triplet correctness prediction with softmax answer classification:

Linear: $y = \sigma(\mathbf{W}\mathbf{x}_{iqa} + b)$
Bilinear: $y = \sigma(\mathbf{x}_{iq}^\top \mathbf{W}\mathbf{x}_a + b)$
MLP: $y = \sigma(\mathbf{W}_2 \max(0, \mathbf{W}_1\mathbf{x}_{iqa}) + b)$

Dataset	Model	Softmax	Binary
Visual7W	Linear	41.6	44.7
	Bilinear	–	63.6
	MLP	50.2	**67.1**
VQA	MLP	61.1	**64.9**

7. Conclusion

Modeling the answer helps for VQA multiple-choice
Are the evaluation protocols of current VQA tasks adequate?
Are current, more complex VQA models learning what we think they are?

For more information, contact {ajabri,ajoulin,lvdmaaten}@fb.com.