Ask Your Neurons: A Neural-based Approach to Answering Questions about Images

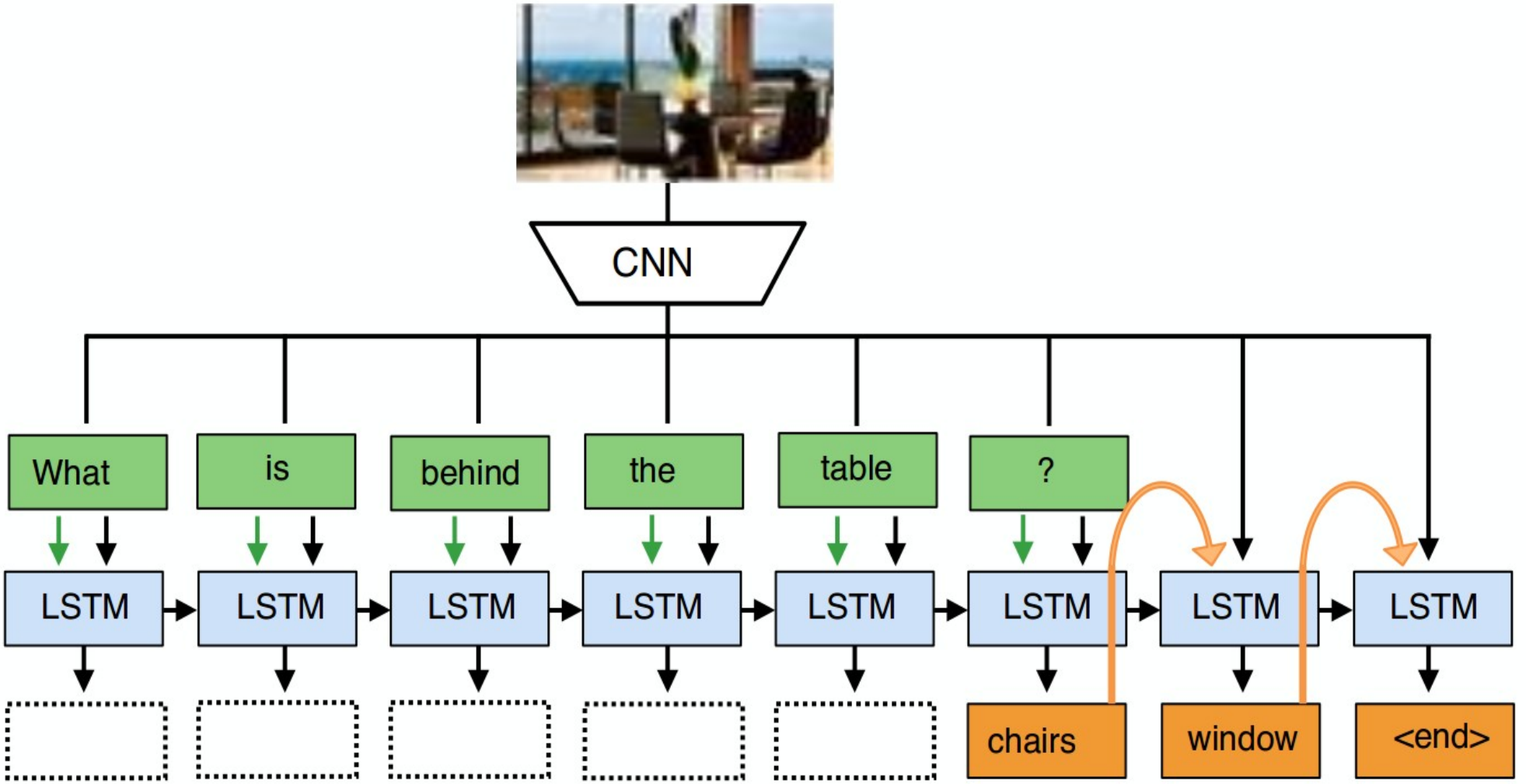
Authors
Malinowski et al.

Experiment by
Huihuang Zheng

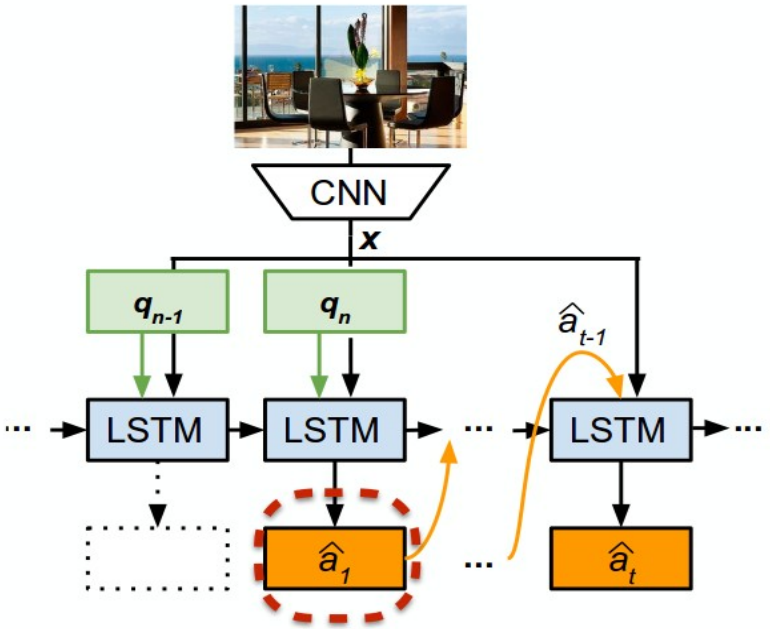
Outline

- Method in the paper
- Successful & fail examples
- Question without image
- Ask questions not relating to image
- Image with Gaussian Filter Smoothing
- Conclusion

Neural Approach to Answer



Formula:



- Predicting answer sequence
  - Recursive formulation

$$\hat{\boldsymbol{a}}_t = \arg\max_{\boldsymbol{a}\in\mathcal{V}} p(\boldsymbol{a}|\boldsymbol{x}, \boldsymbol{q}, \hat{A}_{t-1}; \boldsymbol{\theta}),$$ $\boldsymbol{x}$ - image representation

$$\boldsymbol{q} = [\boldsymbol{q}_1, \ldots, \boldsymbol{q}_{n-1}, [\![?]\!]],$$ $\boldsymbol{q}_j$ - question word index

$\mathcal{V}$ - vocabulary, $\hat{A}_{t-1} = \{\hat{\boldsymbol{a}}_1, \ldots, \hat{\boldsymbol{a}}_{\boldsymbol{t-1}}\}$ - previous answer words

Slide Credit: Malinowski, Rohrbach and Fritz

Experiment 1: Success & Fail

- Experiment setting:
  - Hardware
    - CPU: Intel® Core™ i7-4720HQ CPU @ 2.60GHz × 8
    - GPU: Geforce GTX 960M in laptop
    - Memory: 16GB
  - Tool:
    - Caffe-recurrent (caffe branch with RNN)
    - Python script
      - From original paper example and I modified to do other things

Experiment 1: Success & Fail

- Experiment setting:
  - Dataset:
    - DAQUER test set (dataset in paper)
    - Other dataset:
      - VQA (selected indoor images)*
      - MIT indoor scene
    - Images from Google image search

What Question: Correct

- what is on the bed ?
- clothes

How Many Question: Correct

- how many doors are there?
- 1

What Question: error

- what are the things on the shelf ? (answer: books)
- Book, photo, toy

What Question: error

- what is the colour of the screen ? (answer: red)
- green

How Many Question: error

- how many glass cups are there? (answer 4)
- 1

Summary 1

- The NLP for questions
  - Answer object for what question
  - Answer number for how many question
    - Among first 1000 questions, only 2 (Q574, Q779) answer number fro what question. 1 (Q947) answer color for object question.
- Combining image feature and NLP is not easy.
  - GoogleNet, extract great image feature.

NLP error: Answer number for color

- what color are the bookshelves in this picture in the image84 ? (answer: brown)
- 1

NLP error: Answer Number for What Question

- what objects are on the cabinet below the medal rack? (answer: paper, photo, cup)
- 1

NLP error: Answer object for color

- what color are the drawer knobs ? (answer: white)
- pillow

Experiment 2: Question without image

- Experiment setting:
  - Input black image, DAQUAR questions
  - Compare answers to original answers

- Result:
  - Only 1457 / 5673 (25.68%) answers are different
  - Suggest that LSTM doesn't learn a lot from images in training.

Black vs Origin

- Q0: what is on the left side of the white oven on the floor and on right side of the blue armchair ?
  - (garbage bin)
- Black image: pillow
- True image: pillow

Black vs Origin

- Q1: what is the object placed in front of the filling shelves ?
  - (table)
- Black image: projector_screen
- True image: paper

Black vs Origin

- Q2: how many garbage bins are there ? (2)
- A2:
- Black image: 1
- True image: 1

Experiment 3: Ask questions not relating to image

- Experiment setting:
  - Shuffle DAQUAR test images
  - Compare answers to original answers
- Result:
  - Only 699 / 5673 (12.32%) answers are different

Black vs Shuffle vs Origin

- Q0: what is on the left side of the white oven on the floor and on right side of the blue armchair ?
  - (garbage bin)
- Black: pillow
- Wrong Image: pillow
- True Image: pillow

Black vs Shuffle vs Origin

- Q1: what is the object placed in front of the filling shelves ?
  - (table)
- Black image: projector_screen
- Wrong image: projector_screen
- True image: paper

Black vs Shuffle vs Origin

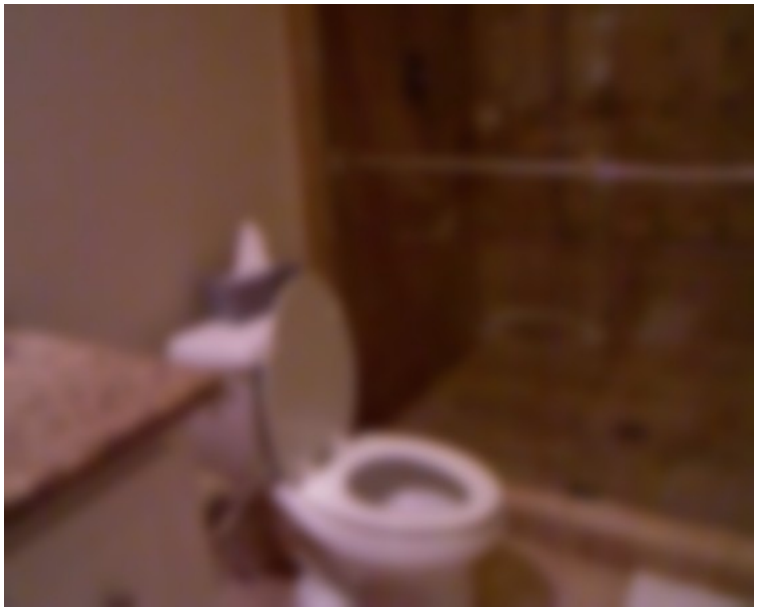
- Q2: how many garbage bins are there ? (2)
- Black image: 1
- Wrong image: 1
- True image: 1

Experiment 3

- Experiment setting:
  - Perform transformation in images to see whether the answer is invariant under some translation
  - Gaussian Smooth DAQUAR test images
    - For 3 color channels, Gaussian filter with size 5.
  - Move down the image top edge by 80 pixels. (origin 425 * 500 pixels)
- Result:
  - 785 / 5673 (13.84% vs 12.32% of random shuffle) answers are different in Gaussian Filter
    - Not invariant to Gaussian Filter
  - 337 / 5673 (5.9 %) answers are different in moving down
    - Somehow invariant to moving, shifting

Wrong Image but Correct Answer!

- Question: what is on the left side of the toilet? (tissue_roll)
- Black & Gaussian: glass
- Wrong & true image: tissue_roll



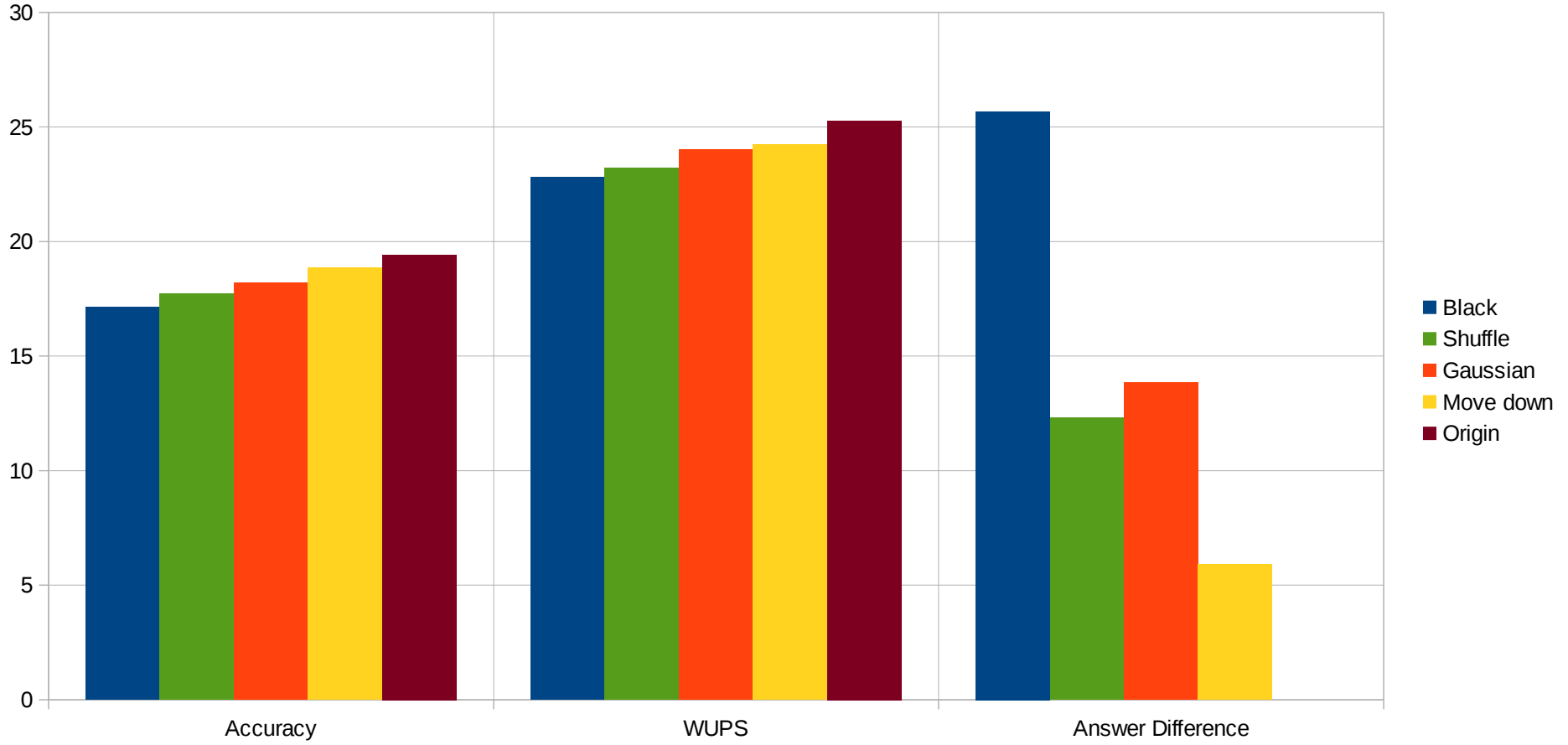
Wrong Image but Correct Answer!

- what is the white object on top of the brown table? (printer)
- Black & Gaussian: vase
- True image: book
- Wrong image: printer

Quantity Summary:

	Accuracy (%)	WUPS (%)	Answer Diff (%)
Black	17.15	22.8	25.68
Random Shuffle	17.73	23.2	12.32
Gaussian Filtered	18.18	24.01	13.83
Movedown	18.86	24.23	5.9
Origin	19.42	25.18	0

Experiment Accuracy (%)



- Accuracy is similar!

Summary 2

- LSTM doesn't learn a lot from image features of CNN in training, which causes bad performance.
- That illustrates why vision + language doesn't improve much from language only

Methods	Accuracy	WUPS @0.9
Baseline: Symbolic (NIPS'14)	7.86%	11.86%
Language Only (Our)	17.15%	22.80%
Vision + Language (Our)	**19.43%**	**25.28%**
Human performance (NIPS'14)	50.20%	50.82%

Result Table from: Malinowski, Rohrbach and Fritz

Conclusion

- In the experiments
  - CNN + LSTM learns great in NLP question
    - Answer numbers for how many question
    - Answer object for what question
    - Answer colors for color question
  - LSTM doesn't learn a lot from CNN feature
    - The result using correct or wrong input images doesn't influence much in result.
    - Still a long way to go.
- Thank you : -)

Reference

- Ask Your Neurons: A Neural-based Approach to Answering Questions about Images.
  - M. Malinowski, M. Rohrbach, and M. Fritz. arXiv 2015. (ICCV'15, Oral, to appear).

- Their ICCV 2015 (Oral) slide

-