Designing deep architectures for Visual Question Answering

Matthieu Cord
Sorbonne University
valeo.ai research lab. Paris

Thanks to H. Ben-younes, R. Cadène
Visual Question Answering

Question Answering:
What does Claudia do?
Visual Question Answering: What does Claudia do?
Visual Question Answering

What does Claudia do?

+ Sitting at the bottom
  Standing at the back...

Visual Question Answering:
Visual Question Answering:

What does Claudia do?

Sitting at the bottom
Standing at the back
...

Solving this task interesting for:
- Study of deep learning models in a multimodal context
- Improving human-machine interaction
- One step to build visual assistant for blind people
Outline

1. Multimodal embedding
   • Deep nets to align text+image
   • learning

2. VQA framework
   • Task modeling
   • Fusion in VQA
   • Reasoning in VQA
Deep semantic-visual embedding

Images

ConvNet

Multimodal space

RNN

Texts

A cat wearing a suit with a red tie and white shirt.
Deep semantic-visual embedding

Image captioning

(CNN, LSTM, ...)

ConvNet

Multimodal space

Semantics of distance
Retrieval by NN search

RNN

Images

Texts

A cat wearing a suit with a red tie and white shirt.

(VAE, GAN, ...)

Image generation
Deep semantic-visual embedding

2D Semantic visual space example:
• Distance in the space has a semantic interpretation
• Retrieval is done by finding nearest neighbors
Deep semantic-visual embedding

- Designing image and text embedding architectures
- Learning scheme for these deep hybrid nets
Deep semantic-visual embedding


Finding beans in burgers: Deep semantic-visual embedding with localization, M. Engilberge et al, CVPR 2018

Visual pipeline:
- ResNet-152 pretrained
- Weldon spatial pooling
- Affine projection

Textual pipeline:
- Pretrained word embedding
- Simple Recurrent Unit (SRU)
- Normalization

\(\theta_0, \theta_1, \theta_2\) and \(\phi\) are the trained parameters
Deep semantic-visual embedding

Finding beans in burgers: Deep semantic-visual embedding with localization, M. Engilberge et al, CVPR 2018

Visual pipeline:
- ResNet-152 pretrained
- Weldon spatial pooling

Textual pipeline:
- Pretrained word embedding
- Simple Recurrent Unit (SRU)

(a, man, in, ski, gear, skiing, on, snow)

Deep semantic-visual embedding

(b) 

Finding beans in burgers: Deep semantic-visual embedding with localization, M. Engilberge et al, CVPR 2018

Visual pipeline:
- ResNet-152 pretrained
- Weldon spatial pooling

Textual pipeline:
- Pretrained word embedding
- Simple Recurrent Unit (SRU)

(a, man, in, ski, gear, skiing, on, snow)
Deep semantic-visual embedding

Finding beans in burgers: Deep semantic-visual embedding with localization, M. Engilberge et al, CVPR 2018

Visual pipeline:
- ResNet-152 pretrained
- Weldon spatial pooling

Textual pipeline:
- Pretrained word embedding
- Simple Recurrent Unit (SRU)

\[ \theta_0 \rightarrow \theta_1 \rightarrow \theta_2 \rightarrow \phi \rightarrow \langle X, V \rangle \]

\( \theta_0, \theta_2 \) and \( \phi \rightarrow \text{Learning using a training set} \)
How to get large training datasets?

Cooking recipes: easy to get large multimodal datasets with aligned data


Deep semantic-visual embedding

Demo Visiir.lip6.fr
## Cross-modal retrieval

<table>
<thead>
<tr>
<th>Query</th>
<th>Closest elements</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A plane in a cloudy sky</strong></td>
<td><img src="image1.png" alt="Image" /> <img src="image2.png" alt="Image" /> <img src="image3.png" alt="Image" /> <img src="image4.png" alt="Image" /> <img src="image5.png" alt="Image" /></td>
</tr>
<tr>
<td><strong>A dog playing with a frisbee</strong></td>
<td><img src="image6.png" alt="Image" /> <img src="image7.png" alt="Image" /> <img src="image8.png" alt="Image" /> <img src="image9.png" alt="Image" /> <img src="image10.png" alt="Image" /></td>
</tr>
<tr>
<td><strong>A herd of sheep standing on top of snow covered field.</strong></td>
<td><img src="image11.png" alt="Image" /> <img src="image12.png" alt="Image" /> <img src="image13.png" alt="Image" /> <img src="image14.png" alt="Image" /> <img src="image15.png" alt="Image" /></td>
</tr>
<tr>
<td><strong>There are sheep standing in the grass near a fence.</strong></td>
<td><img src="image16.png" alt="Image" /> <img src="image17.png" alt="Image" /> <img src="image18.png" alt="Image" /> <img src="image19.png" alt="Image" /> <img src="image20.png" alt="Image" /></td>
</tr>
<tr>
<td><strong>some black and white sheep a fence dirt and grass</strong></td>
<td><img src="image21.png" alt="Image" /> <img src="image22.png" alt="Image" /> <img src="image23.png" alt="Image" /> <img src="image24.png" alt="Image" /> <img src="image25.png" alt="Image" /></td>
</tr>
</tbody>
</table>
Cross-modal retrieval and localization

Visual grounding examples:

• Generating multiple heat maps with different textual queries

Finding beans in burgers: Deep semantic-visual embedding with localization, M. Engilberge et al, CVPR 2018
Cross-modal retrieval and localization

Emergence of color understanding:
Outline

1. Multimodal embedding
   • Deep nets to align text+image
   • Learning

2. Visual Question Answering
   • Task modeling
   • Fusion in VQA
   • Reasoning in VQA
VQA
Visual Question Answering

Images

Questions
What color is the cat’s tie?

Multimodal space

Answers
red

What does the man ride while wearing a black wet suit?
Ground truth: surfboard
IMG+BOW: jacket (0.35)
2-VIS+LSTM: surfboard (0.53)
BOW: tie (0.30)

DAQUAR 2136
What is right of table?
Ground truth: shelves
IMG+BOW: shelves (0.33)
2-VIS+BLSTM: shelves (0.28)
LSTM: shelves (0.20)

Does it appear to be rainy?
Does this person have 20/20 vision?

How many slices of pizza are there?
Is this a vegetarian pizza?
What color is the fire hydrant on the left?

Green
VQA

What color is the fire hydrant on the right?

Yellow
Who is wearing glasses?

Different answers

Similar images

man

woman

⇒ Need very good Visual and Question (deep) representations
⇒ Full scene understanding
⇒ Need High level multimodal interaction modeling
⇒ Merging operators, attention and reasoning

@VQA workshop, CVPR 2017
Vanilla VQA scheme: 2 deep + fusion

Question Representation

Image
Question: Is the lady with the blue fur wearing glasses?

VQA: the output space

Yes
VQA: the output space

VQA Dataset [Antol et al. 2015]
- released for the VQA Challenge Workshop at CVPR 2016
- Each pair (image, question) is associated with 10 correct answers

Figure: Example of an (image, question, answers) triplet from VQA dataset
VQA: the output space

Evaluation metric

\[
acc_{vqa}(\text{answer}) = \min \left( 1, \frac{\# \text{ humans that provided that answer}}{3} \right)
\]

Volumes:

- Train set: 82,783 images, 248,349 questions and answers
- Val set: 40,504 images, 121,512 questions and answers
- Test set: 81,434 images, 244,302 questions

Output space representation:
=> Classify over the most frequent answers (3000/95%)
Question: Is the lady with the blue fur wearing glasses?
VQA processing

**Image**
- Convolutional Network (VGG, ResNet,....)
- Detection system (EdgeBoxes, Faster-RCNN, ...)

**Question**
- Bag-of-words
- Recurrent Network (RNN, LSTM, GRU, SRU, ...)

**Learning**
- Fixed answer vocabulary
- Classification (cross-entropy)
Fusion in VQA
VQA: fusion

Concatenation & projection: \( y = W \begin{bmatrix} q \\ v \end{bmatrix} \)

Element-wise sum: \( y = (Wq) + (Vv) \)

Element-wise product: \( y = (Wq) \odot (Vv) \)

Multi-layer perceptron: \( y = MLP \begin{bmatrix} q \\ v \end{bmatrix} \)

Is the lady with the purple fur wearing glasses? Yes

\[ y \] Softmax \[ \text{no} \]
VQA: fusion

Is the lady with the purple fur wearing glasses?

Concatenation & projection: \( y = W \begin{bmatrix} q \\ v \end{bmatrix} \)

Element-wise sum: \( y = (Wq) + (Vv) \)

Element-wise product: \( y = (Wq) \odot (Vv) \)

Multi-layer perceptron: \( y = MLP \left( \begin{bmatrix} q \\ v \end{bmatrix} \right) \)
VQA: bilinear fusion

[Kim, Jin-Hwa et al. Hadamard Product for Low-rank Bilinear Pooling, ICLR 2017]

Is the lady with the purple fur wearing glasses?

Bilinear model:
score for class $k = \text{bilinear combination of dimensions in } q \text{ and } v$

$$y^k = \sum_{i=1}^{d_q} \sum_{j=1}^{d_v} T^{ijk} q^i v^j$$

$$y = \mathcal{T} \times_1 q \times_2 v$$
VQA: bilinear fusion

\[ y^k = \sum_{i=1}^{d_q} \sum_{j=1}^{d_v} T^{ijk} q^i v^j \]

Learn the 3-ways tensor coeff.
- Different than the Signal Proc. Tensor analysis (representation)

Problem: \( q, v \) and \( y \) are of dimension \( \sim 2000 \)
- \( => 8 \text{ billion free parameters} \) in the tensor

Need to reduce the tensor size:
- Idea: structure the tensor to reduce the number of parameters
VQA: bilinear fusion

Tensor structure:

Tucker decomposition:

\[ \mathcal{T} = ((\mathcal{T}_c \times_1 W_q) \times_2 W_v) \times_3 W_o \]

\( \Leftrightarrow \) constrain the rank of each unfolding of \( \mathcal{T} \)
VQA: bilinear fusion
VQA: bilinear fusion
VQA: bilinear fusion

\[(\mathcal{T}_c \times_1 (q^\top W_q)) \times_2 (v^\top W_v)\]
VQA: bilinear fusion

\[ y = ((\mathbf{T}_c \times_1 (q^T \mathbf{W}_q)) \times_2 (v^T \mathbf{W}_v)) \times_3 \mathbf{W}_o \]
VQA: bilinear fusion

Other ways of structuring the tensor of parameters

Compact Bilinear Pooling (MCB)

Low-Rank Bilinear Pooling (MLB)

Tucker Decomposition (MUTAN)

VQA: bilinear fusion [AAAI 2019]

BLOCK [Ben-Younes et al. 2019], extension with a structured $\mathcal{T}$ using a block-term decomposition [De Lathauwer 2008]:

$$
\mathcal{T} := \sum_{r=1}^{R} \mathcal{D}_r \times_1 A_r \times_2 B_r \times_3 C_r
$$

$\mathcal{D}_r \in \mathbb{R}^{L \times M \times N}$, $A_r \in \mathbb{R}^{d_q \times L}$, $B_r \in \mathbb{R}^{d_v \times M}$ and $C_r \in \mathbb{R}^{d_o \times N}$
VQA: BLOCK fusion [AAAI 2019]
Classical attention architecture:

Comparing fusion schemes on the VQA Dataset 2.0

| Description                          | Reference          | $|\Theta|$ | All   | Yes/no | Number | Other  |
|--------------------------------------|--------------------|----------|-------|--------|--------|--------|
| (1) Linear                           | Sum                | 8M       | 58.48 | 71.89  | 36.56  | 52.09  |
| (2) Non-linear                        | Concat MLP         | 13M      | 63.85 | 81.34  | 43.75  | 53.48  |
| (3) B + count-sketching               | MCB (Fukui et al. 2016) | 32M     | 61.23 | 79.73  | 39.13  | 50.45  |
| (4) B + Tucker decomp.                | Tucker (Ben-Younes et al. 2017) | 14M   | 64.21 | 81.81  | 42.28  | 54.17  |
| (5) B + CP decomp.                    | MLB (Kim et al. 2017) | 16M   | 64.88 | 81.34  | 43.75  | 53.48  |
| (6) B + low-rank on the 3rd mode slices | MFB (Yu et al. 2017a) | 24M   | 65.56 | 82.35  | 41.54  | 56.74  |
| (7) Combination of (4) and (6)        | MUTAN (Ben-Younes et al. 2017) | 14M   | 65.19 | 82.22  | 42.1   | 55.94  |
| (8) Higher order fusion               | MFH (Yu et al. 2018) | 48M   | 65.72 | 82.82  | 40.39  | 56.94  |
| (9) B + Block-term decomposition      | BLOCK              | 18M     | 66.41 | 82.86  | 44.76  | 57.3   |
VQA: bilinear fusion

Multiple ways of learning a merging function between two vector spaces

- Linear projections
- Deep fusions
- Bilinear models, simplified by:
  - sketching techniques,
  - tensor decompositions framework
- higher-order fusion
Q: Are there an equal number of large things and metal spheres?
What is reasoning (for VQA)?

Attentional reasoning

Relational reasoning

Iterative reasoning

Compositional reasoning
VQA: reasoning

What is reasoning (for VQA)?

**Attentional reasoning**: given a certain context (i.e. Q), focus only on the relevant subparts of the image.

**Relational reasoning**

**Iterative reasoning**

**Compositional reasoning**
VQA: attentional reasoning

Idea: focusing only on parts of the image relevant to Q

- Each region scored according to the question

What is sitting on the desk in front of the boys?

- Representation = sum of all (weighted) embeddings
What is sitting on the desk in front of the boys?
VQA: attentional reasoning

What is sitting on the desk in front of the boys?
VQA: attentional reasoning

Attentional glimpse in most of recent strategies [MLB, MCB, MUTAN…]
VQA: attentional reasoning

What is sitting on the desk in front of the boys?

What are on the shelves in the background?

Tucker Decomposition with Structured Sparsity

Laptops

Books
VQA: attentional reasoning

Focusing on multiple regions: Multi-glimpse attention

Where is the smoke coming from?
VQA: attentional reasoning with Multi-glimpse attention

Where is the smoke coming from?

Focus on the train

Focus on the smoke

Attention mechanism

\[ \sum p_i^1 v_i \]

\[ \sum p_i^2 v_i \]

“train”
VQA: attentional reasoning with Multi-glimpse attention

(a) Question: Where is the woman? - Answer: on the elephant

(b) Question: Where is the smoke coming from? - Answer: train
VQA: attentional reasoning

Evaluation on VQA dataset:
Best MUTAN score of 67.36% on test-std

Human performances about 83% on this dataset

The winner of the VQA Challenge in CVPR 2017 (and CVPR 2018) integrates adaptive grid selection from additional region detection learning process
VQA: attentional reasoning

Underlying reasoning hypothesis: answering a question requires information about objects and their attributes. **Important:** for each region, only its intermediate representation is used.
VQA: reasoning

What is reasoning (for VQA)?

**Attentional reasoning**: given a certain context (i.e. Q), focus only on the relevant subparts of the image

**Relational reasoning**: object detection + mutual relationships (spatial, semantic,...), merging both with Q

**Iterative reasoning**

**Compositional reasoning**
Determine the answer using relevant objects and relationships

**Question:** Are both men wearing ties?

**Answer:** No
VQA: reasoning

What is reasoning (for VQA)?

**Attentional reasoning**: given a certain context (i.e. Q), focus only on the relevant subparts of the image.

**Relational reasoning**: object detection + mutual relationships (spatial, semantic,...), merging both with Q.

**Iterative reasoning**: refining the attention step-by-step, each time extracting a different piece of information from the image.
Iterative Reasoning

At least 3 elementary steps are required to answer the question

- Find bicycles
- Find the bicycle that has a basket
- Find what is in this basket

Stacked attention: iteratively refining visual attention and question representation

What are sitting in the basket on a bicycle?

Zichao Yang et. al., *Stacked Attention Networks for Image Question Answering*, CVPR 2016
Multimodal Relational Reasoning for VQA

MUREL system:
- Vector representation for Attention process
- Spatial and semantic contexts to model relations between image regions
- Iterative process / Multistep reasoning

Cadene et al., MuRel: Multimodal Relational Reasoning for Visual Question Answering CVPR 2019
MuRel: Multimodal Relational Reasoning for VQA
MuRel: Multimodal Relational Reasoning for VQA

Original Image

Step #1

Step #2

Step #3

What game are they playing on the Wii?

bowling✓

What position is the front dog in?

sitting✓

What is the man holding?

kite ✓

What color is the jacket on the right?

green x
MuRel: Multimodal Relational Reasoning for VQA

Step #1

What is on the top of her head?

Step #2

Is she wearing a ring?

Step #3

What is she eating?

hat

yes

donut
### VQA v2.0 dataset

<table>
<thead>
<tr>
<th>Model</th>
<th>Yes/No</th>
<th>test-dev Num.</th>
<th>Other</th>
<th>All</th>
<th>test-dev</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bottom-up</td>
<td>81.82</td>
<td>44.21</td>
<td>56.05</td>
<td>65.32</td>
<td>65.67</td>
<td></td>
</tr>
<tr>
<td>Graph Att.</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>66.18</td>
<td></td>
</tr>
<tr>
<td>MUTAN†</td>
<td>82.88</td>
<td>44.54</td>
<td>56.50</td>
<td>66.01</td>
<td>66.38</td>
<td></td>
</tr>
<tr>
<td>MLB†</td>
<td>83.58</td>
<td>44.92</td>
<td>56.34</td>
<td>66.27</td>
<td>66.62</td>
<td></td>
</tr>
<tr>
<td>DA-NTN</td>
<td>84.29</td>
<td>47.14</td>
<td>57.92</td>
<td>67.56</td>
<td>67.94</td>
<td></td>
</tr>
<tr>
<td>Pythia</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>68.05</td>
<td>-</td>
</tr>
<tr>
<td>Counter</td>
<td>83.14</td>
<td><strong>51.62</strong></td>
<td><strong>58.97</strong></td>
<td><strong>68.09</strong></td>
<td><strong>68.41</strong></td>
<td></td>
</tr>
<tr>
<td>MuRel</td>
<td><strong>84.77</strong></td>
<td>49.84</td>
<td>57.85</td>
<td>68.03</td>
<td>68.41</td>
<td></td>
</tr>
</tbody>
</table>

### TDIUC dataset (12 different categories)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Bottom-up</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Scene Reco.</td>
<td>93.96</td>
<td>93.06</td>
<td>93.80</td>
<td>96.11</td>
</tr>
<tr>
<td>Sport Reco.</td>
<td>93.47</td>
<td>92.77</td>
<td>95.55</td>
<td>96.20</td>
</tr>
<tr>
<td>Color Attr.</td>
<td>66.86</td>
<td>68.54</td>
<td>60.16</td>
<td>74.43</td>
</tr>
<tr>
<td>Other Attr.</td>
<td>56.49</td>
<td>56.72</td>
<td>54.36</td>
<td>58.19</td>
</tr>
<tr>
<td>Activity Reco.</td>
<td>51.60</td>
<td>52.35</td>
<td>60.10</td>
<td>63.83</td>
</tr>
<tr>
<td>Pos. Reasoning</td>
<td>35.26</td>
<td>35.40</td>
<td>34.71</td>
<td>41.19</td>
</tr>
<tr>
<td>Object Reco.</td>
<td>86.11</td>
<td>85.54</td>
<td>86.98</td>
<td>89.41</td>
</tr>
<tr>
<td>Absurd</td>
<td>96.08</td>
<td>84.82</td>
<td><strong>100.00</strong></td>
<td>99.8</td>
</tr>
<tr>
<td>Util. and Afford.</td>
<td>31.58</td>
<td><strong>35.09</strong></td>
<td>31.48</td>
<td>21.43</td>
</tr>
<tr>
<td>Object Presence</td>
<td>94.38</td>
<td>93.64</td>
<td>94.55</td>
<td>95.75</td>
</tr>
<tr>
<td>Counting</td>
<td>48.43</td>
<td>51.01</td>
<td>53.25</td>
<td>61.78</td>
</tr>
<tr>
<td>Sentiment</td>
<td>60.09</td>
<td><strong>66.25</strong></td>
<td>64.38</td>
<td>60.65</td>
</tr>
<tr>
<td>Overall (A-MPT)</td>
<td>67.81</td>
<td>67.90</td>
<td>69.11</td>
<td>71.56</td>
</tr>
<tr>
<td>Overall (H-MPT)</td>
<td>59.00</td>
<td><strong>60.47</strong></td>
<td>60.08</td>
<td>59.30</td>
</tr>
<tr>
<td>Overall Accuracy</td>
<td>84.26</td>
<td>81.86</td>
<td>85.03</td>
<td>88.20</td>
</tr>
</tbody>
</table>

**Figure 1:** Examples from our balanced VQA dataset.
Datasets and challenges

Many initiatives to improve datasets and evaluate reasoning as:

**VQA v2.0** [Y. Goyal, D. Batra, D. Parikh, CVPR 2017]

**TDIUC dataset** and challenge (Task Driven Image Understanding Challenge)

**CLEVR dataset** [J. Johnson et al, CVPR 2017]
- Questions about visual reasoning including attribute identification, counting, comparison, spatial relationships, and logical operations.

**GQA dataset** (2019) for compositional Q answering over real-world images
- 22M diverse reasoning questions generated from a scene graph

**Visual dialogue task**: to hold a dialog with humans in natural, conversational language about visual content
MLIA/Chordettes team:
Matthieu Cord  http://webia.lip6.fr/~cord
A. Dapogny (Postdoc), PhD T. Robert, T. Mordan, H. BenYounes, R. Cadene, E. Mehr, M. Engilberge, Y. Chen, A. Saporta, N. Thome (CNAM Pr 10% associate)

CVPR 2019 **MUREL: Multimodal Relational Reasoning for Visual Question Answering**
R. Cadene, H. Ben-younes, N. Thome, M. Cord


ICCV 2017 **MUTAN: Multimodal Tucker Fusion for Visual Question Answering**
H. Ben-Younes*, R. Cadene*, N. Thome, M. Cord

Pytorch code: https://github.com/Cadene