# Designing deep architectures for Visual Question Answering

Matthieu Cord
Sorbonne University
valeo.ai research lab. Paris

Thanks to H. Ben-younes, R. Cadène

SORBONNE UNIVERSITÉ

Valeo
SMART TECHNOLOGY FOR SMARTER CARS

# Visual Question Answering

Question Answering:

What does Claudia do?

# Visual Question Answering

Visual Question Answering:

What does Claudia do?

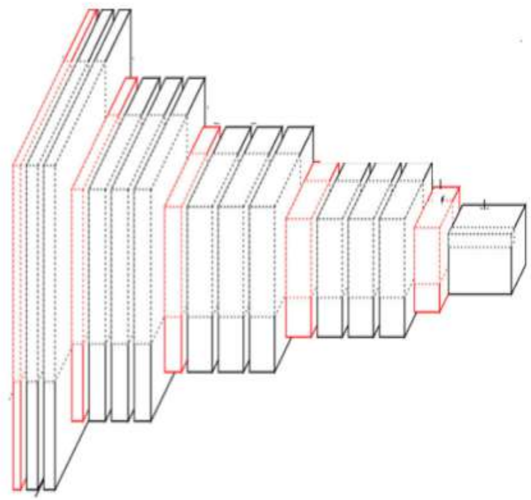
+

# Visual Question Answering

Visual Question Answering:

What does Claudia do?

+



Sitting at the bottom

Standing at the back

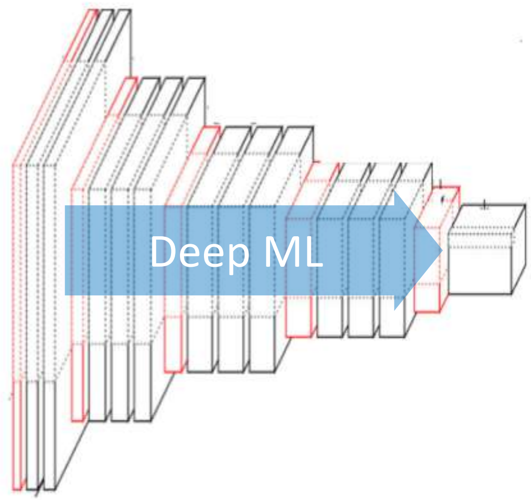
...

# Visual Question Answering

Visual Question Answering:

What does Claudia do?

+




Sitting at the bottom

Standing at the back

...

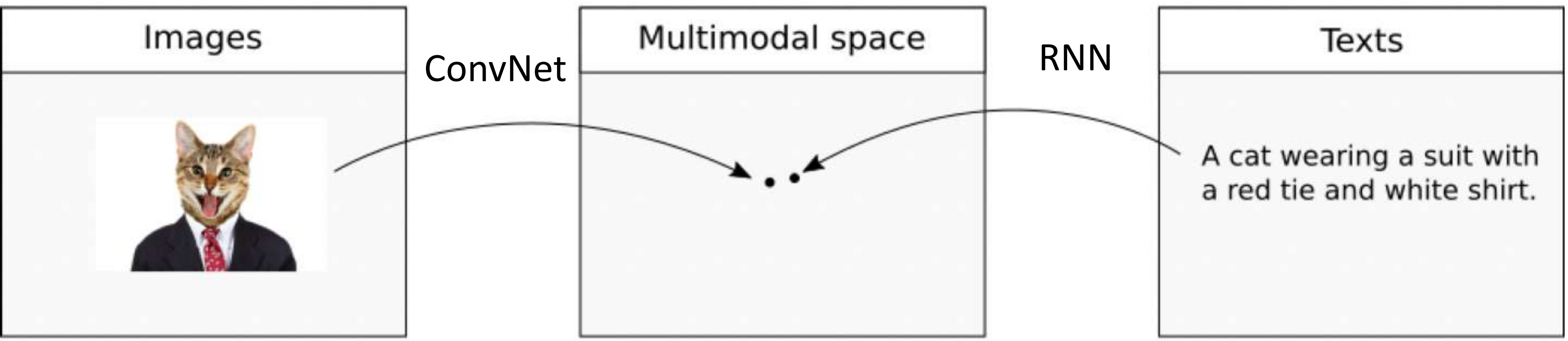
Solving this task interesting for:

- Study of deep learning models in a multimodal context
- Improving human-machine interaction
- One step to build visual assistant for blind people

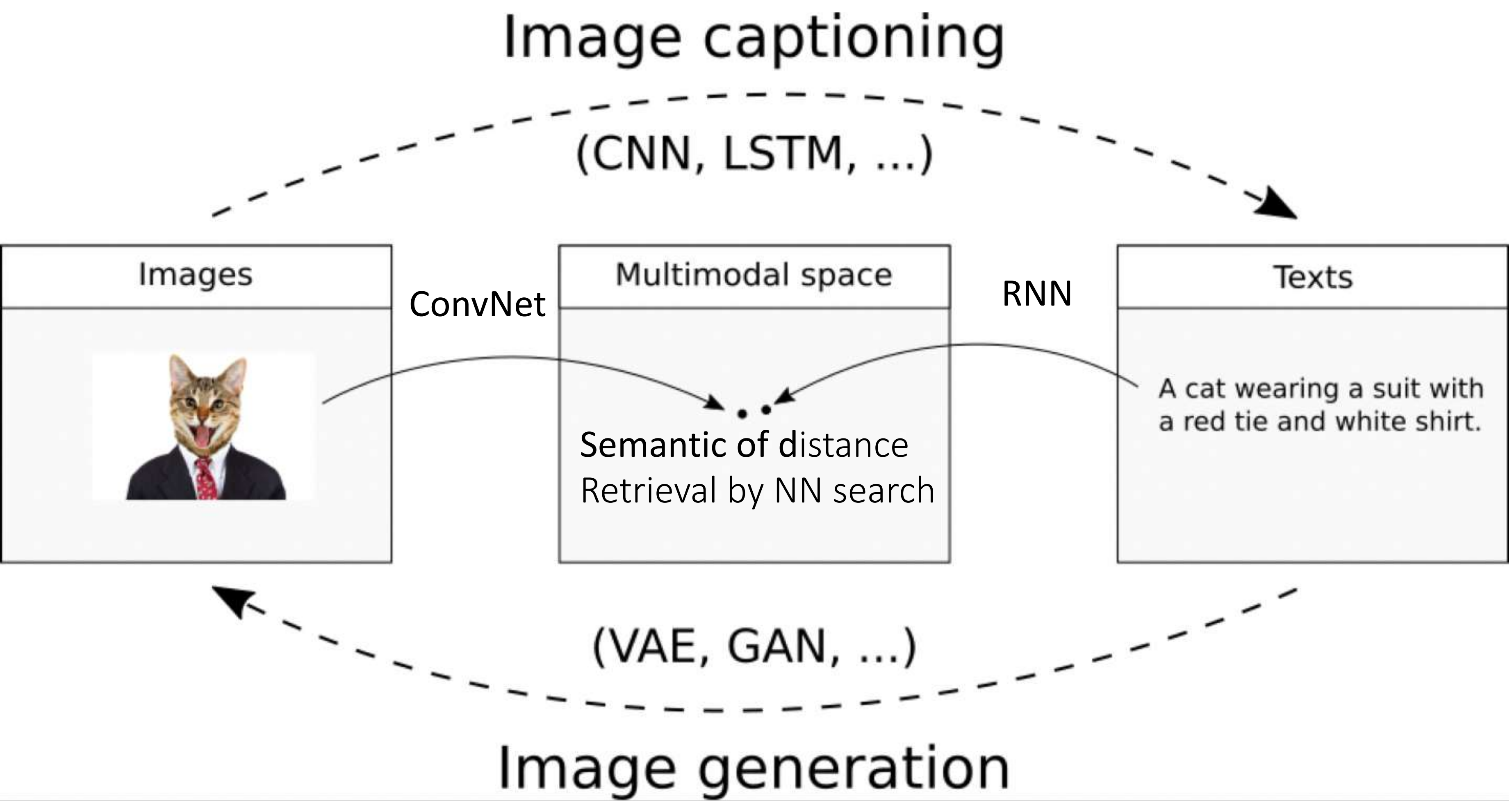
# Outline

1. **Multimodal embedding**
   - **Deep nets to align text+image**
   - **learning**
2. VQA framework
   - Task modeling
   - Fusion in VQA
   - Reasoning in VQA

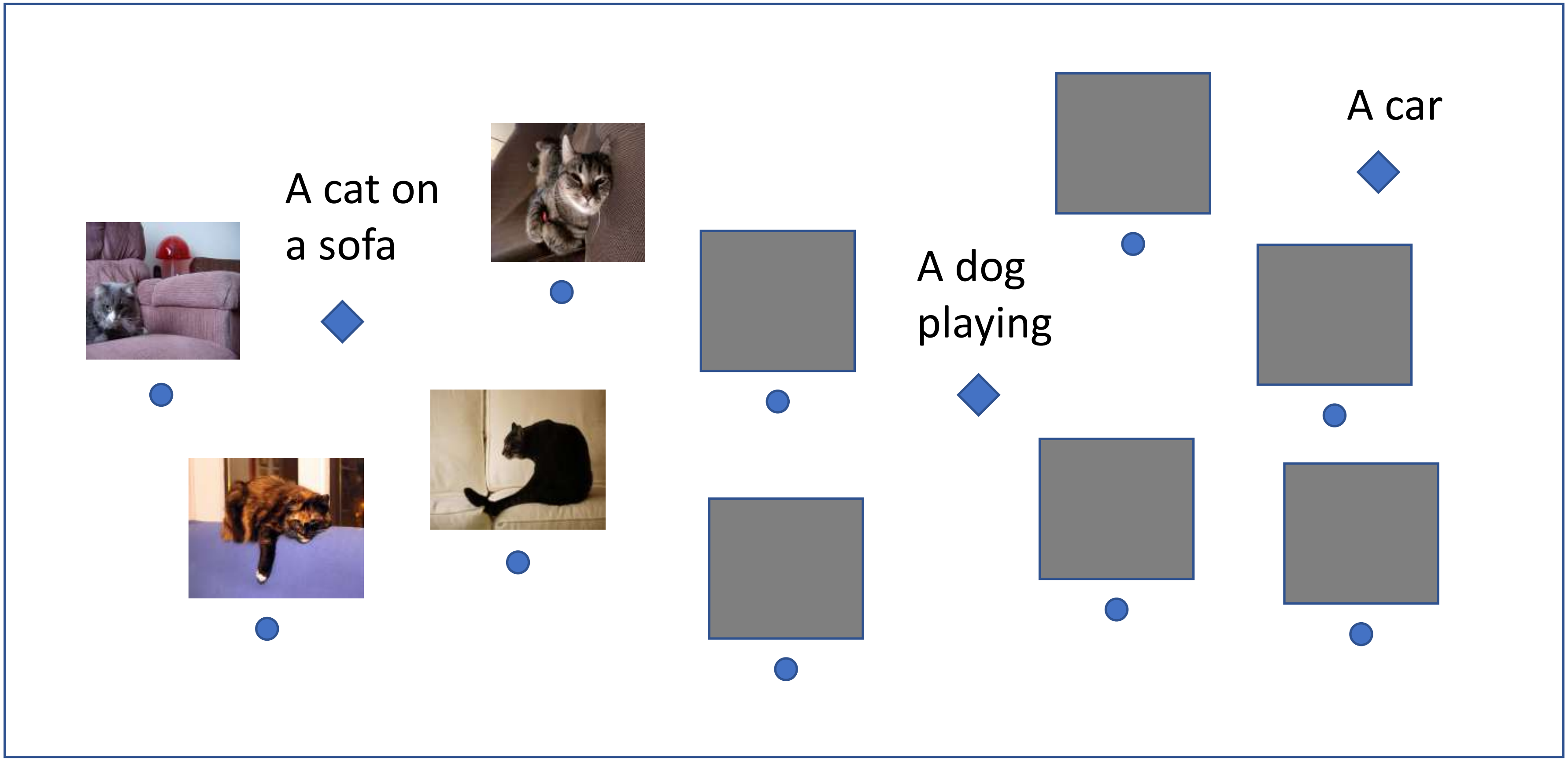
# Deep semantic-visual embedding



# Deep semantic-visual embedding



# Deep semantic-visual embedding

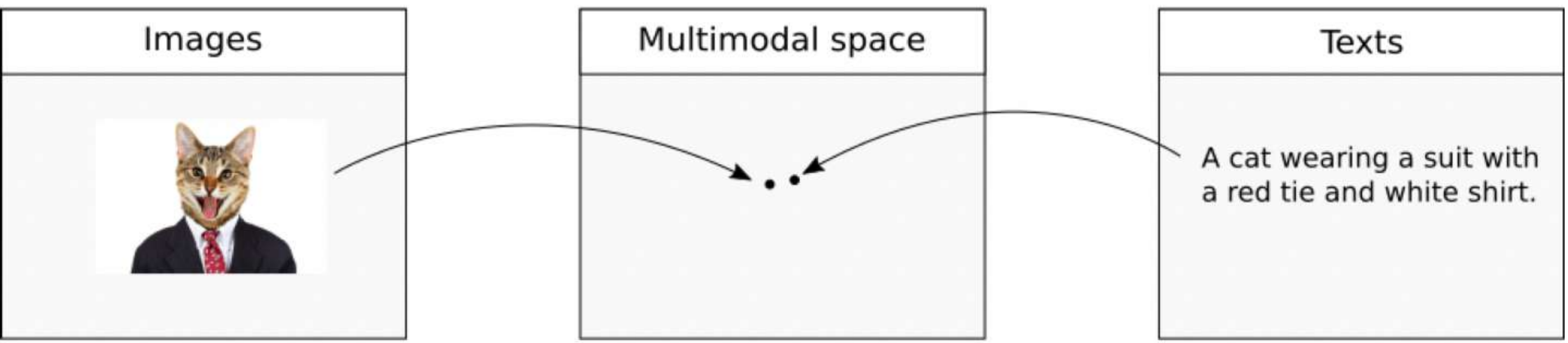



2D Semantic visual space example:

- Distance in the space has a semantic interpretation
- Retrieval is done by finding nearest neighbors

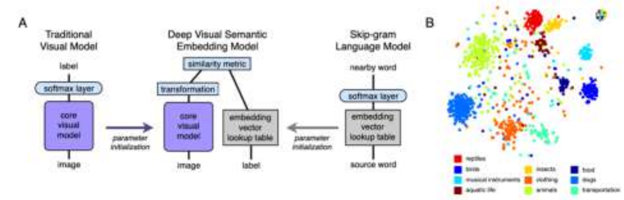
# Deep semantic-visual embedding

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



# Deep semantic-visual embedding

**DeViSE: A Deep Visual-Semantic Embedding Model,**
A. Frome et al, NIPS 2013



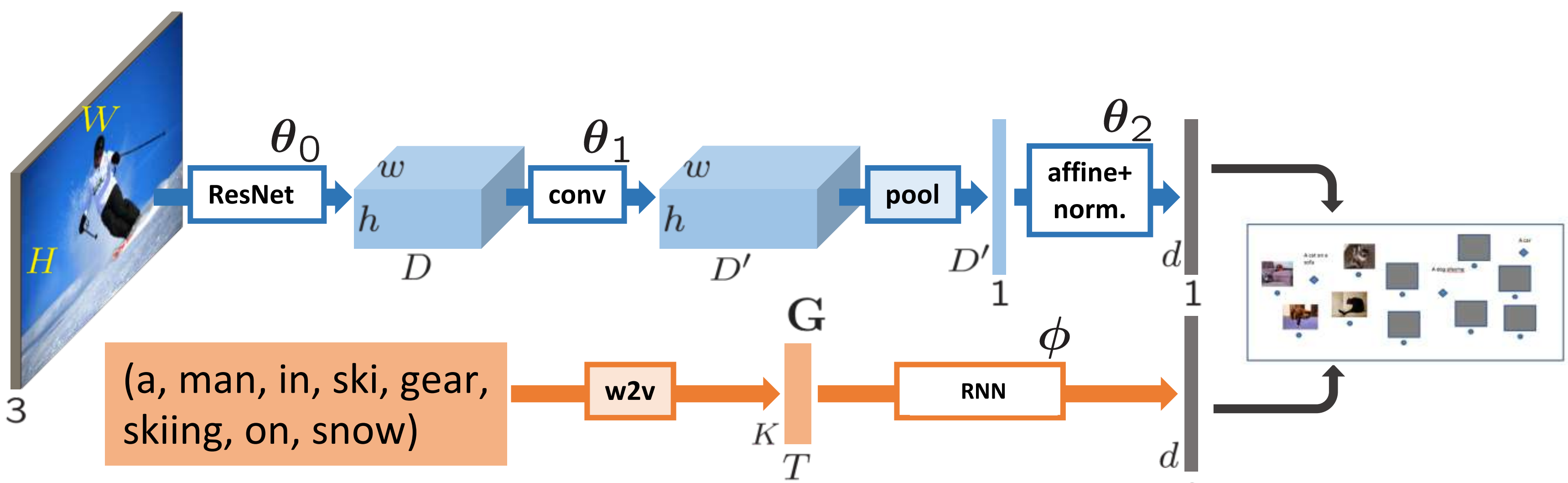

**Finding beans in burgers: Deep semantic-visual embedding with localization,** M. Engilberge et al, CVPR 2018

**Visual pipeline:**

- ResNet-152 pretrained
- Weldon spatial pooling
- Affine projection

**Textual pipeline:**

- Pretrained word embedding
- Simple Recurrent Unit (SRU)
- Normalization




$\theta_{0:2}$ and $\phi$ are the trained parameters

# Deep semantic-visual embedding

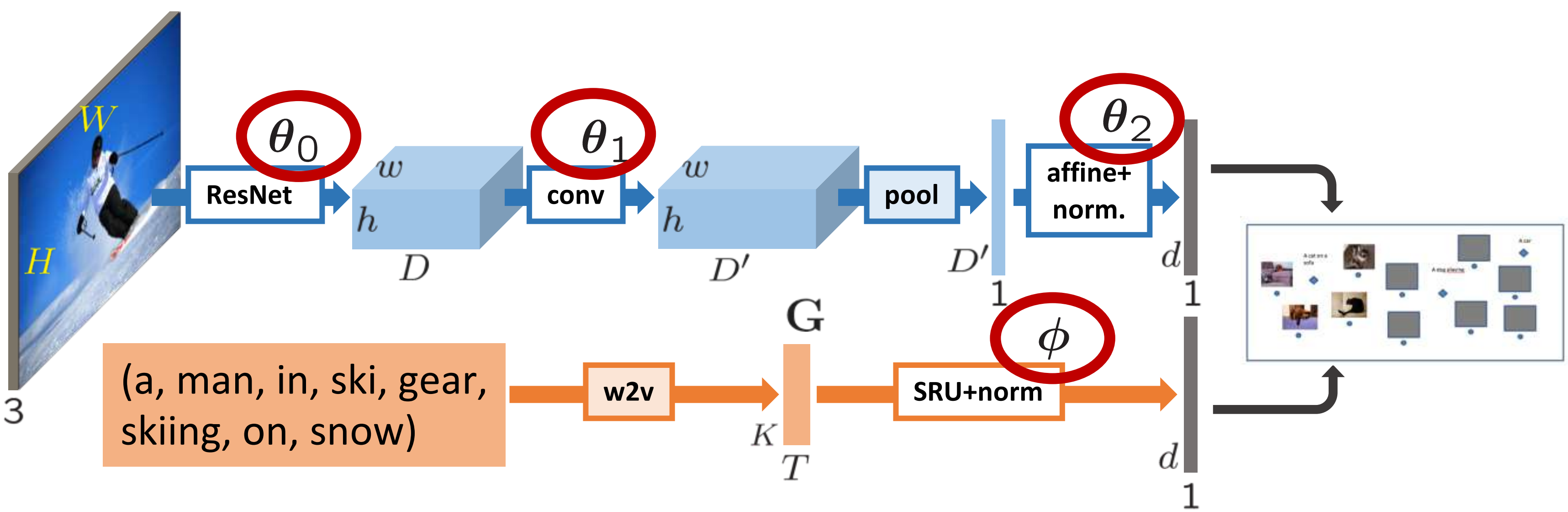
**Finding beans in burgers: Deep semantic-visual embedding with localization,**
M. Engilberge et al, CVPR 2018

**Visual pipeline:**

- ResNet-152 pretrained
- Weldon spatial pooling

**Textual pipeline:**

- Pretrained word embedding
- Simple Recurrent Unit (SRU)



# Deep semantic-visual embedding

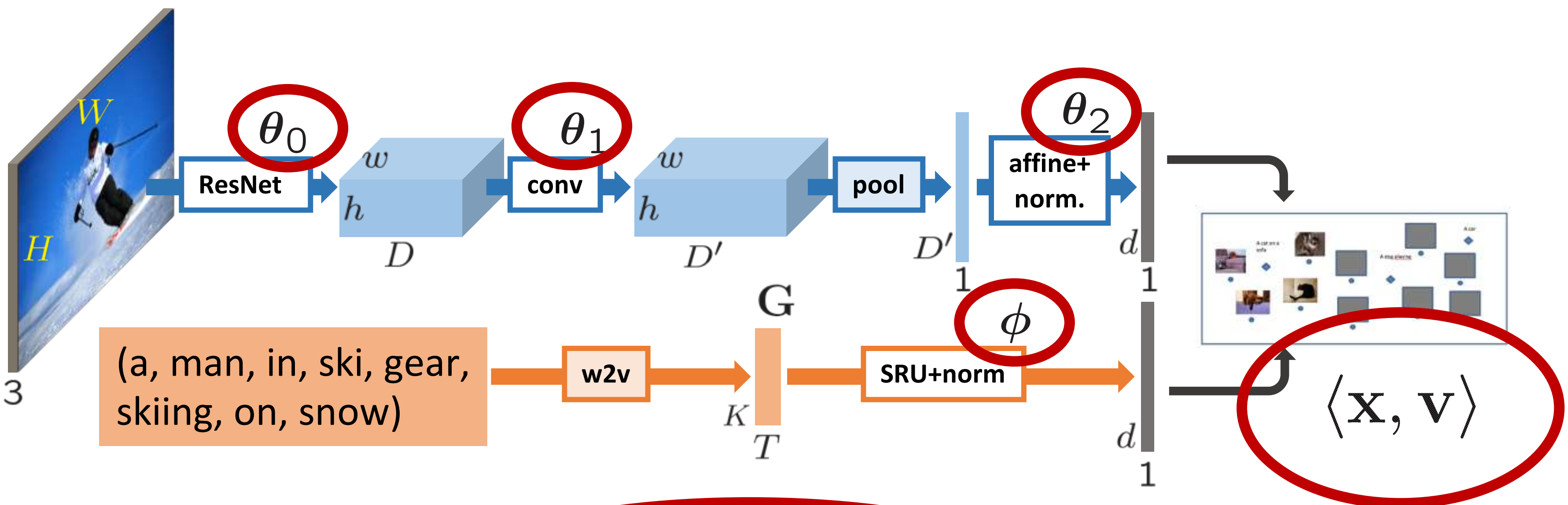
**Finding beans in burgers: Deep semantic-visual embedding with localization,**
M. Engilberge et al, CVPR 2018

**Visual pipeline:**

- ResNet-152 pretrained
- Weldon spatial pooling

**Textual pipeline:**

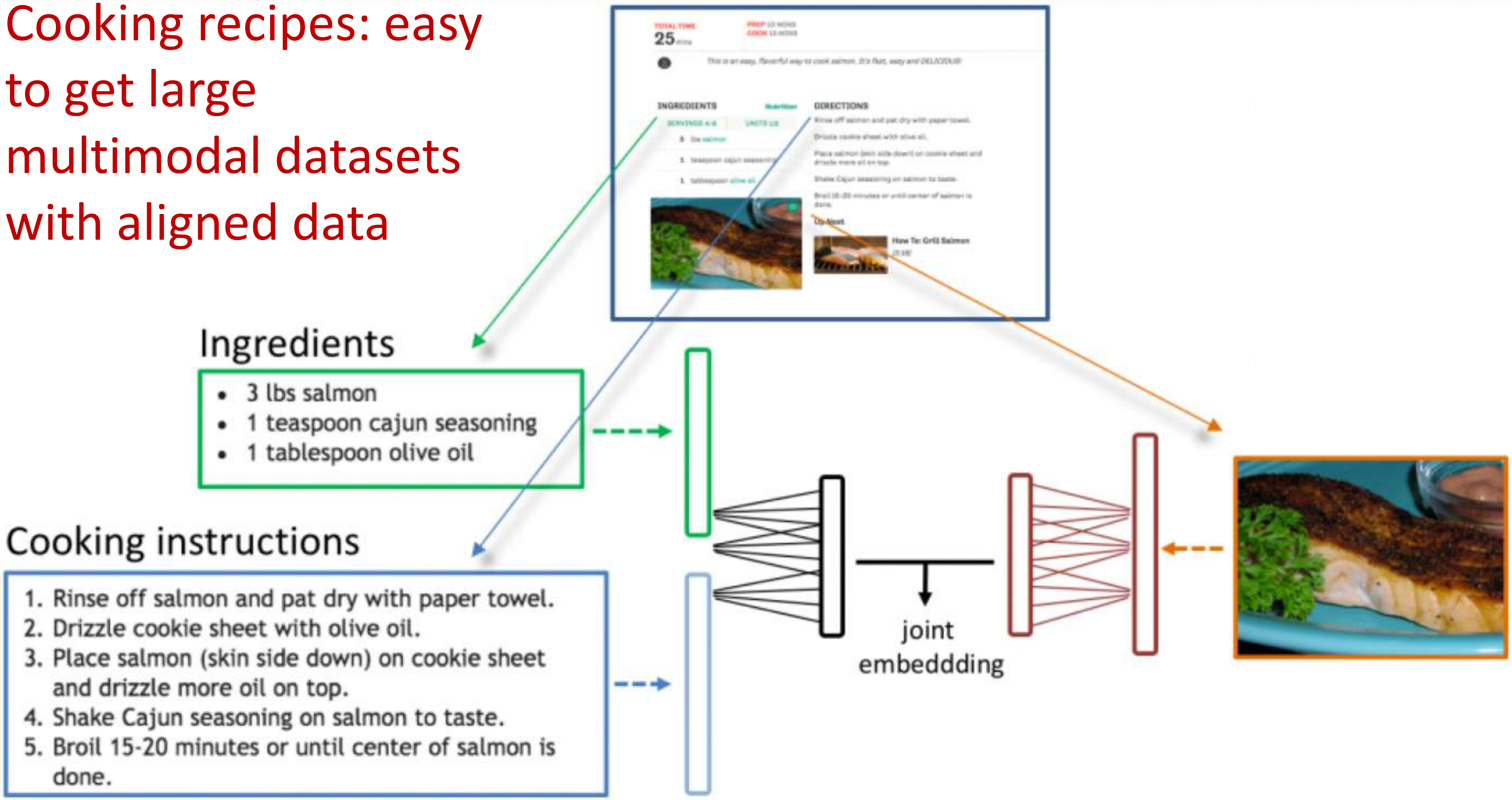
- Pretrained word embedding
- Simple Recurrent Unit (SRU)




$\theta_{0:2}$ and $\phi$ => Learning using a training set

# How to get large training datasets?

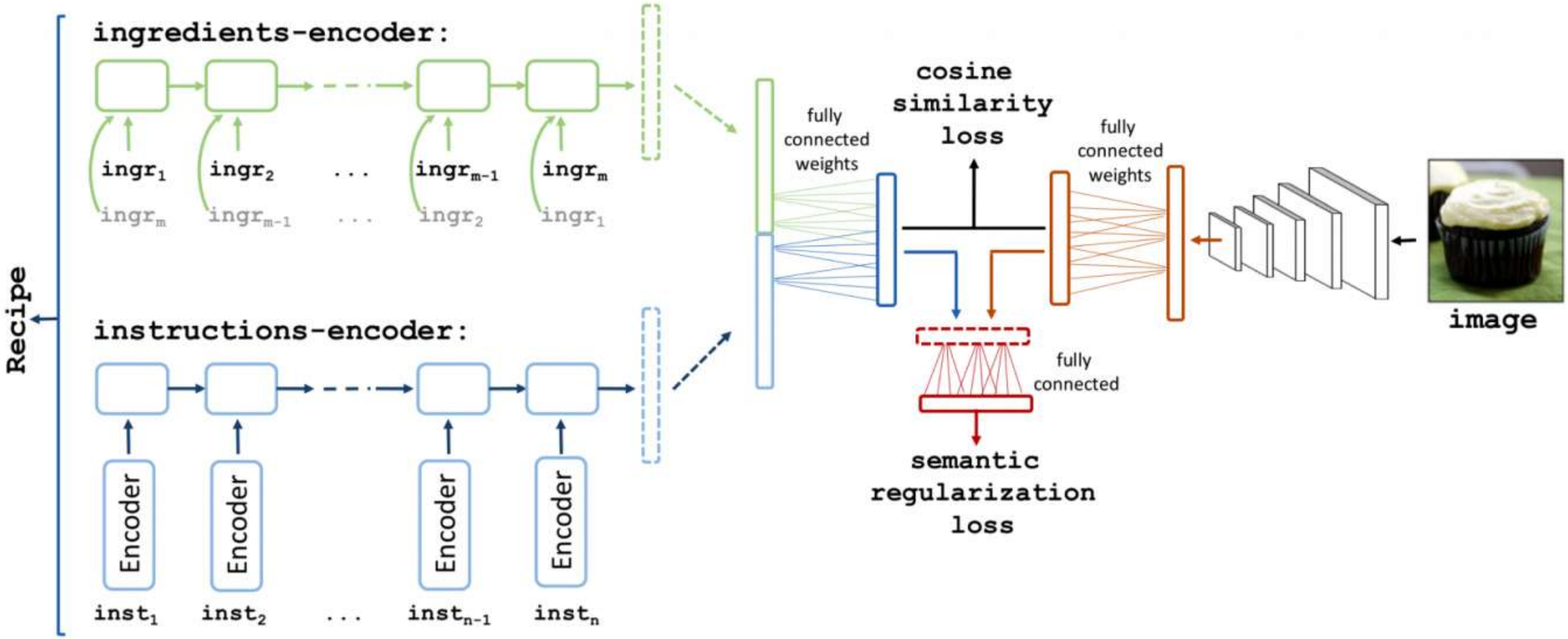
Cooking recipes: easy to get large multimodal datasets with aligned data

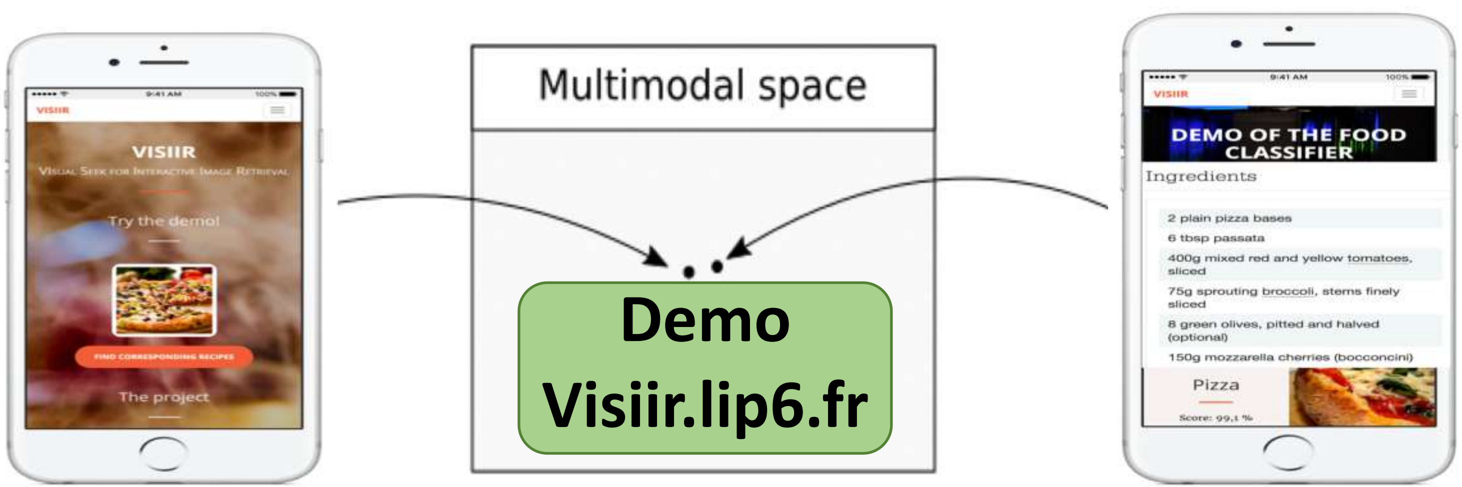



Learning Cross-modal Embeddings for Cooking Recipes and Food Images. A. Salvador, et al. CVPR 2017

[Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings]() M. Carvalho, R. Cadene, D. Picard, L. Soulier, N. Thome, M. Cord, SIGIR (2018)

# Deep semantic-visual embedding



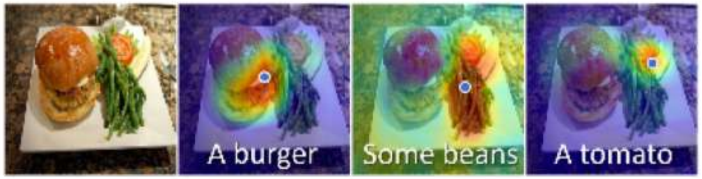



# Cross-modal retrieval

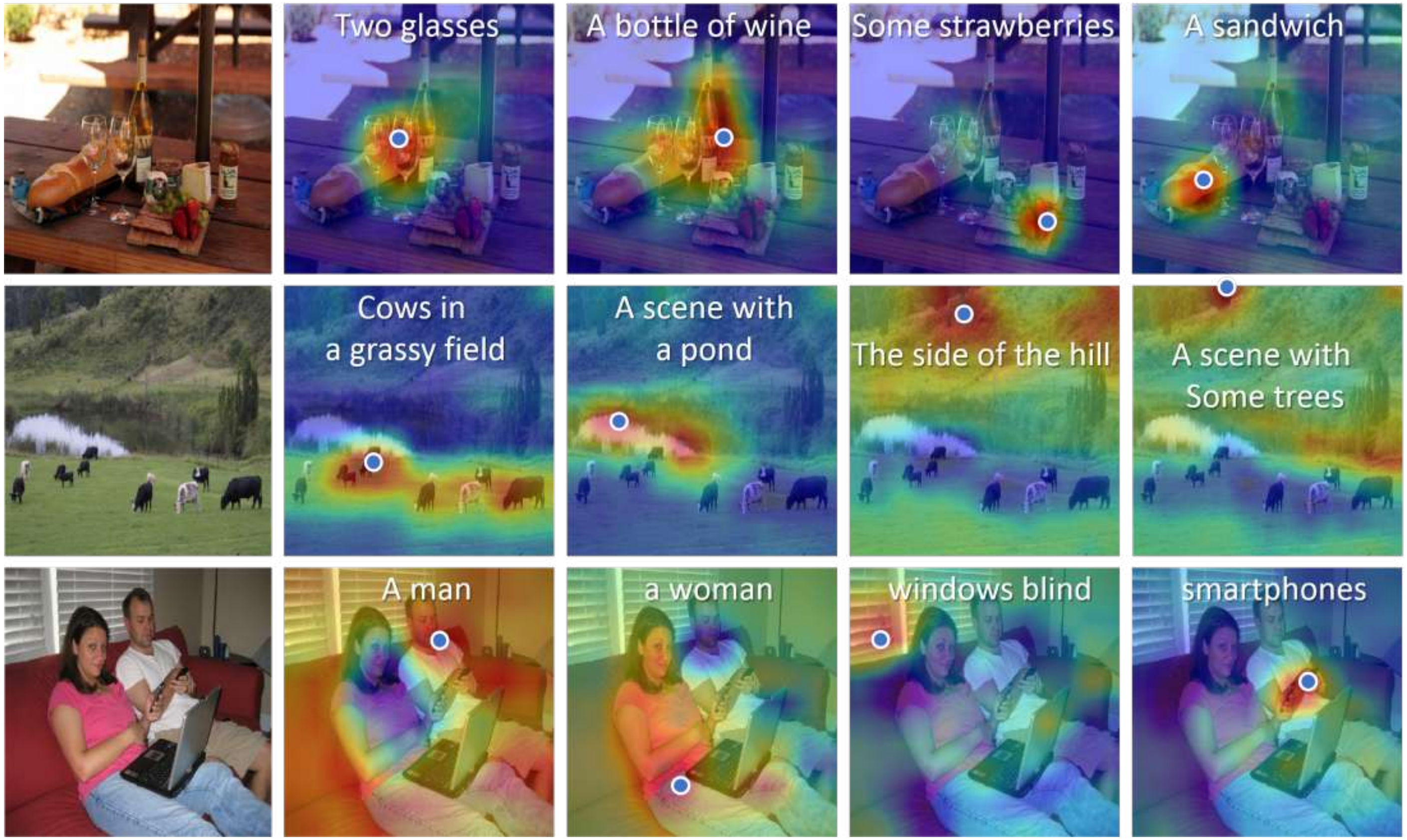
| Query | Closest elements | | | | |
|---|---|---|---|---|---|
| A plane in a cloudy sky |  |  |  |  |  |
| A dog playing with a frisbee |  |  |  |  |  |
|  | 1. **A herd of sheep standing on top of snow covered field.**<br>2. There are sheep standing in the grass near a fence.<br>3. some black and white sheep a fence dirt and grass | | | | |

# 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

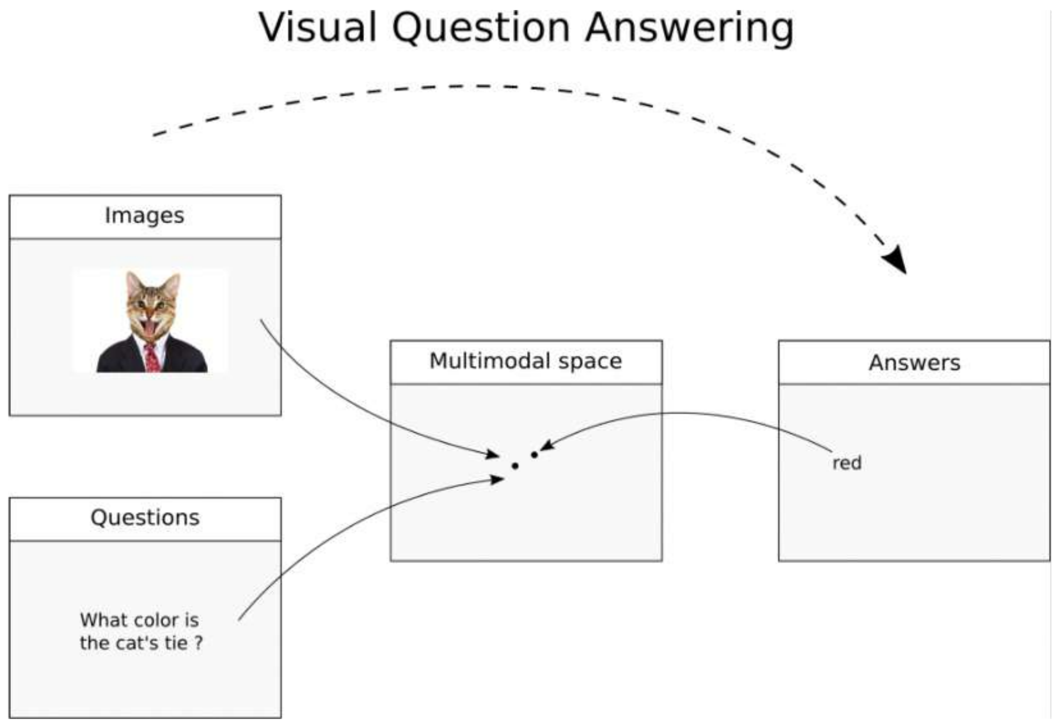
# Cross-modal retrieval and localization

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




COCOQA 15756
**What does the man rid while wear-ing 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?

# VQA

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?**

Similar images

**man** | **woman**

Different answers

@VQA workshop, CVPR 2017

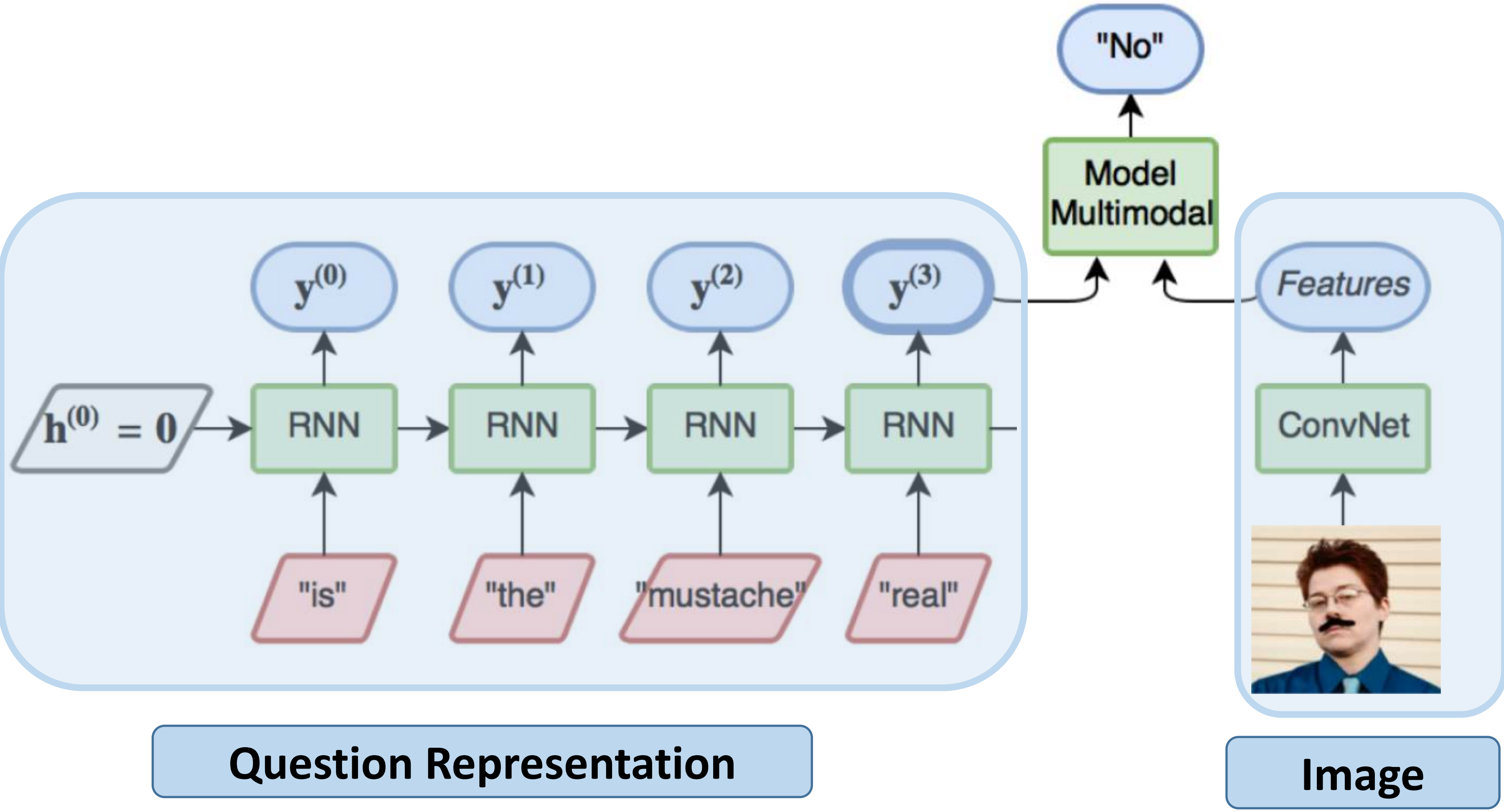
⇒ Need very good Visual and Question (deep) representations

  ⇒ Full scene understanding

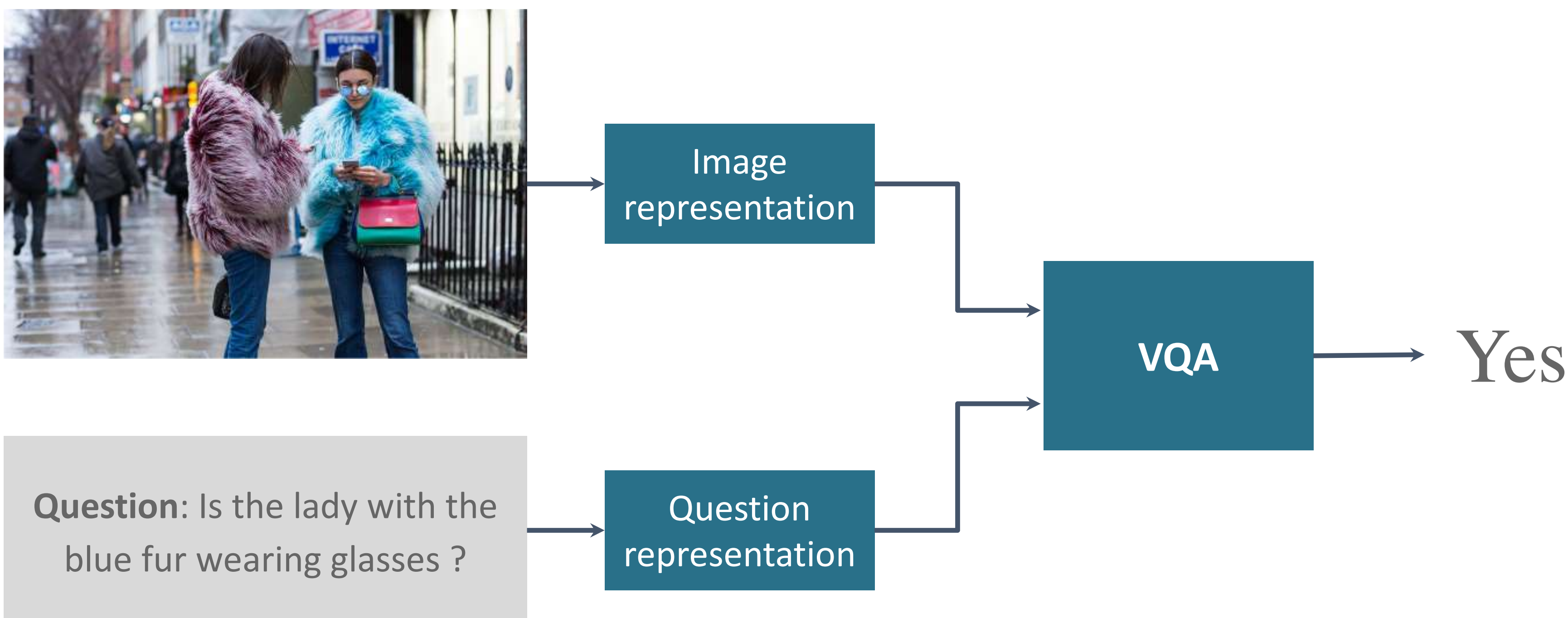
⇒ Need High level multimodal interaction modeling

  ⇒ Merging operators, attention and reasoning

# Vanilla VQA scheme: 2 deep + fusion



# VQA: the output space



# 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


Q: Is road work being performed?

Ground-Truth Answers:

| | |
|---|---|
| (1) yes | (6) yes |
| (2) yes | (7) yes |
| (3) yes | (8) yes |
| (4) yes | (9) yes |
| (5) yes | (10) yes |

Q: What does the sign say to do?

Ground-Truth Answers:

| | |
|---|---|
| (1) stop | (6) stop |
| (2) stop | (7) stop |
| (3) stop | (8) stop |
| (4) stop | (9) stop |
| (5) stop | (10) stop |

Q: What color is the sign?

Ground-Truth Answers:

| | |
|---|---|
| (1) red | (6) red |
| (2) red | (7) red and white |
| (3) red | (8) red |
| (4) red | (9) red/white |
| (5) red | (10) red |

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

# VQA: the output space

Evaluation metric

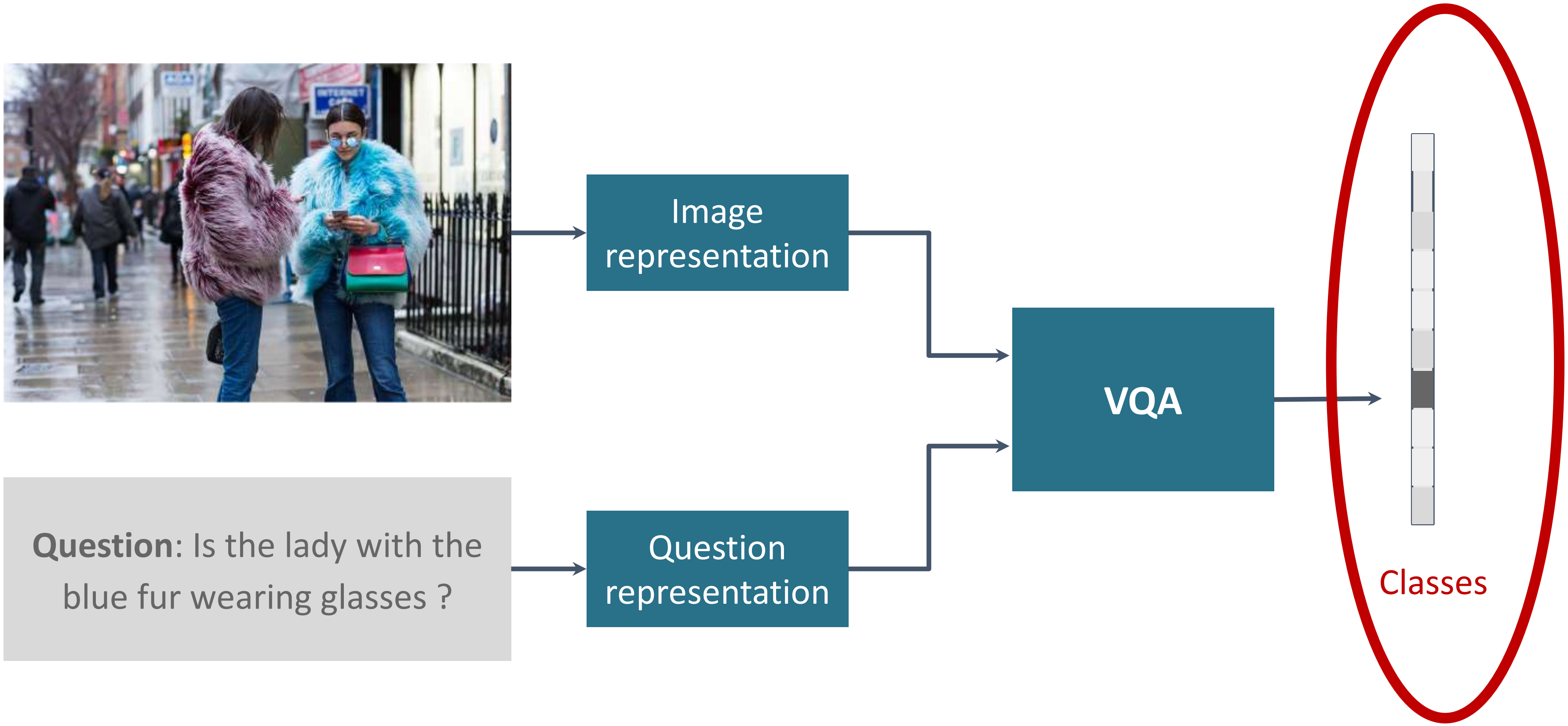
$$acc_{vqa}(answer) = \min\left(1, \frac{\#\text{ humans that provided that answer}}{3}\right) \quad (2)$$

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%)

# VQA: the output space



# VQA processing

**Image**

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

**Multimodal Fusion Reasoning**

**Question**

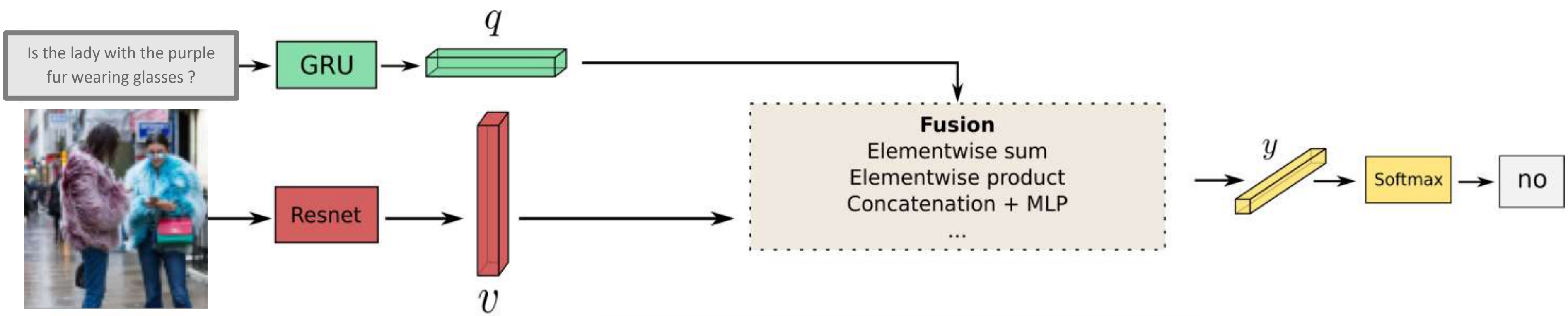
- *Bag-of-words*
- Recurrent Network (RNN, LSTM, GRU, SRU, …)

**Learning**

- Fixed answer vocabulary
- Classification (cross-entropy)

# Fusion in VQA

# VQA: fusion



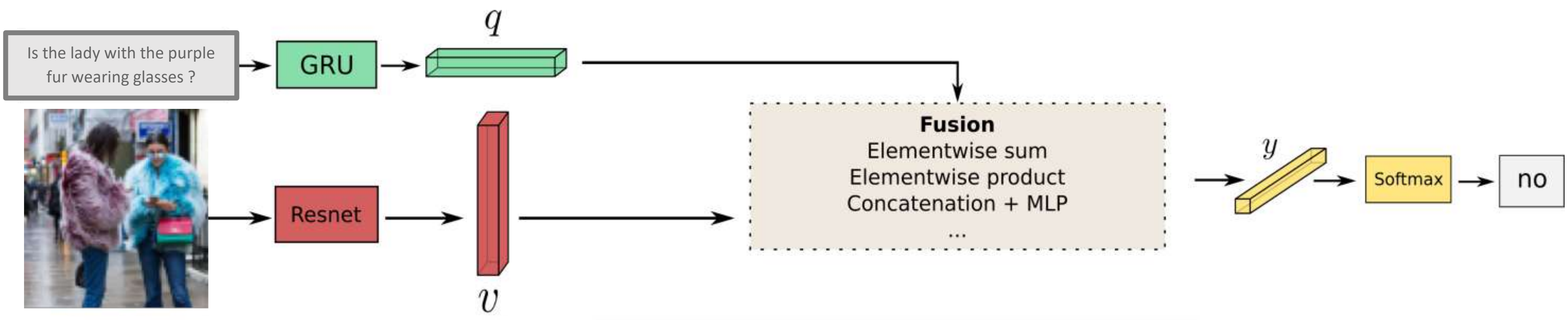

Concatenation & projection : $y = \boldsymbol{W} \begin{bmatrix} \mathbf{q} \\ \mathbf{v} \end{bmatrix}$

Element-wise sum : $y = (\boldsymbol{W}\mathbf{q}) + (\boldsymbol{V}\mathbf{v})$

Element-wise product : $y = (\boldsymbol{W}\mathbf{q}) \odot (\boldsymbol{V}\mathbf{v})$

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

# VQA: fusion




Concatenation & projection : $y = \boldsymbol{W} \begin{bmatrix} \mathbf{q} \\ \mathbf{v} \end{bmatrix}$

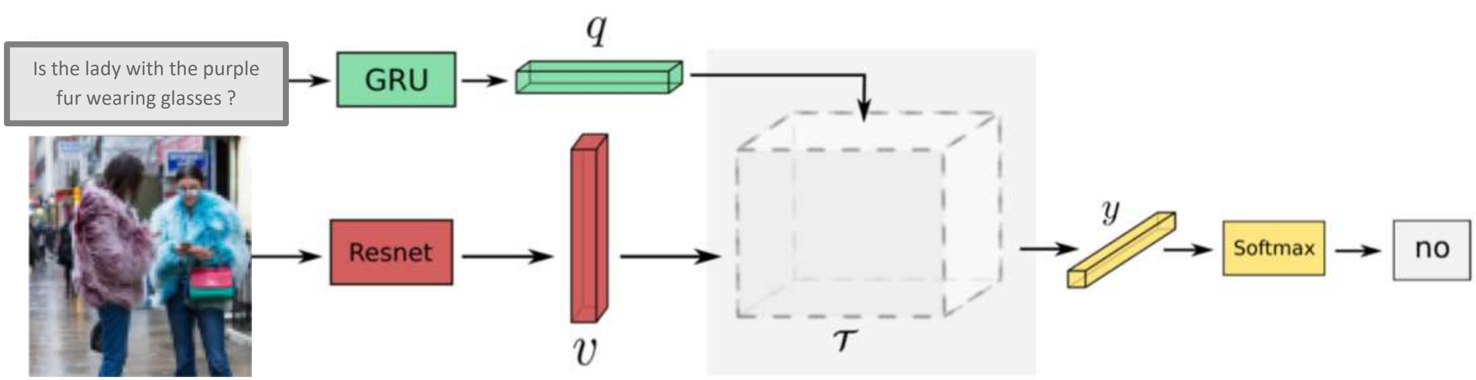
Element-wise sum : $y = (\boldsymbol{W}\mathbf{q}) + (\boldsymbol{V}\mathbf{v})$

Element-wise product : $y = (\boldsymbol{W}\mathbf{q}) \odot (\boldsymbol{V}\mathbf{v})$

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

# VQA: bilinear fusion

[Fukui, Akira et al. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding, CVPR 2016]
[Kim, Jin-Hwa et al. Hadamard Product for Low-rank Bilinear Pooling, ICLR 2017]



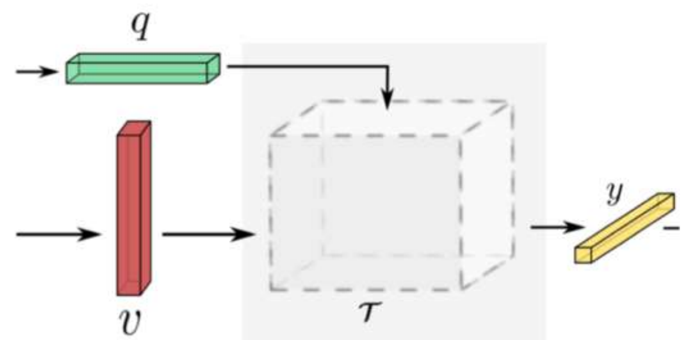

Bilinear model:
score for class k = bilinear combination of dimensions in **q** and **v**

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

$$\mathbf{y} = \mathcal{T} \times_1 \mathbf{q} \times_2 \mathbf{v}$$

# VQA: bilinear fusion

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




Learn the 3-ways tensor coeff.

- Different than the Signal Proc. Tensor analysis (representation)

Problem: **q**, **v** and **y** are of dimension ~ 2000
=> **8 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

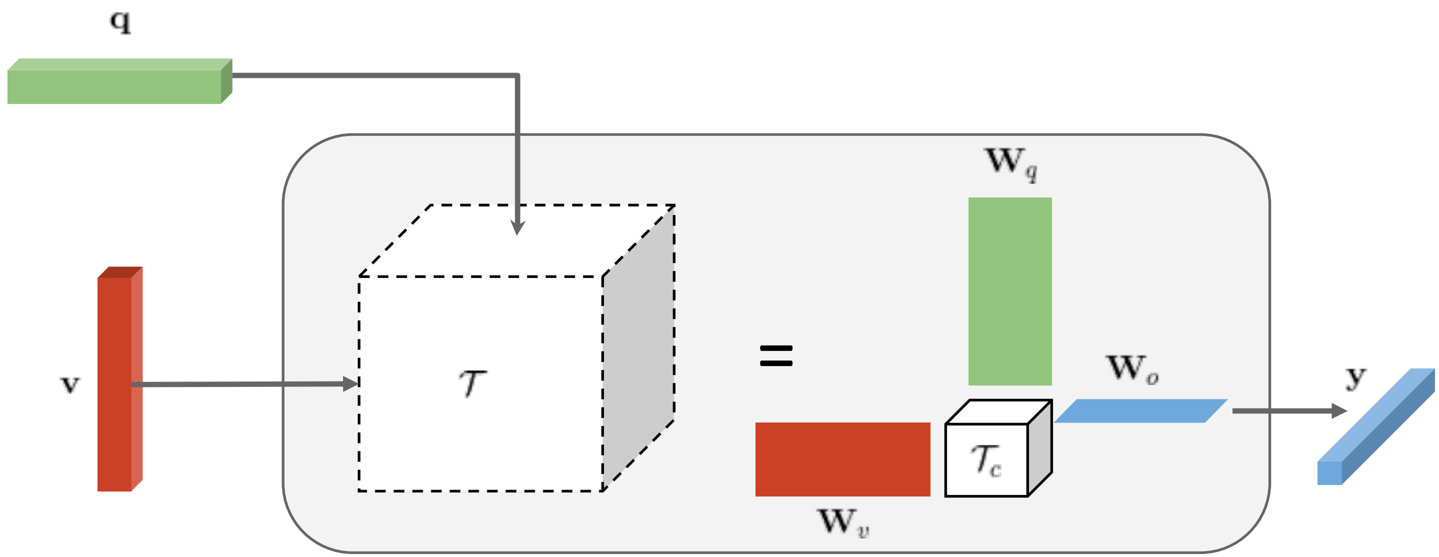
Tensor structure:

Tucker decomposition:

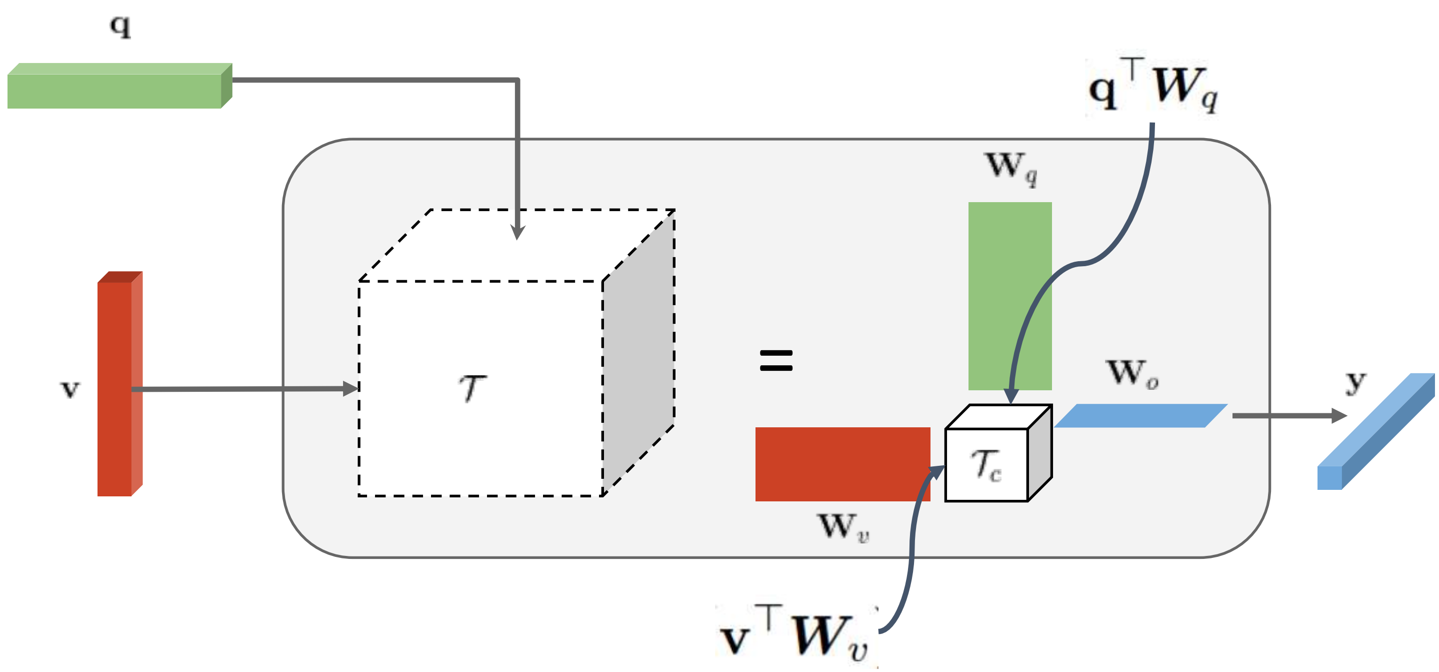
$$\mathcal{T} = ((\mathcal{T}_c \times_1 \mathbf{W}_q) \times_2 \mathbf{W}_v) \times_3 \mathbf{W}_o$$

$\Leftrightarrow$ constrain the rank of each unfolding of $\mathcal{T}$

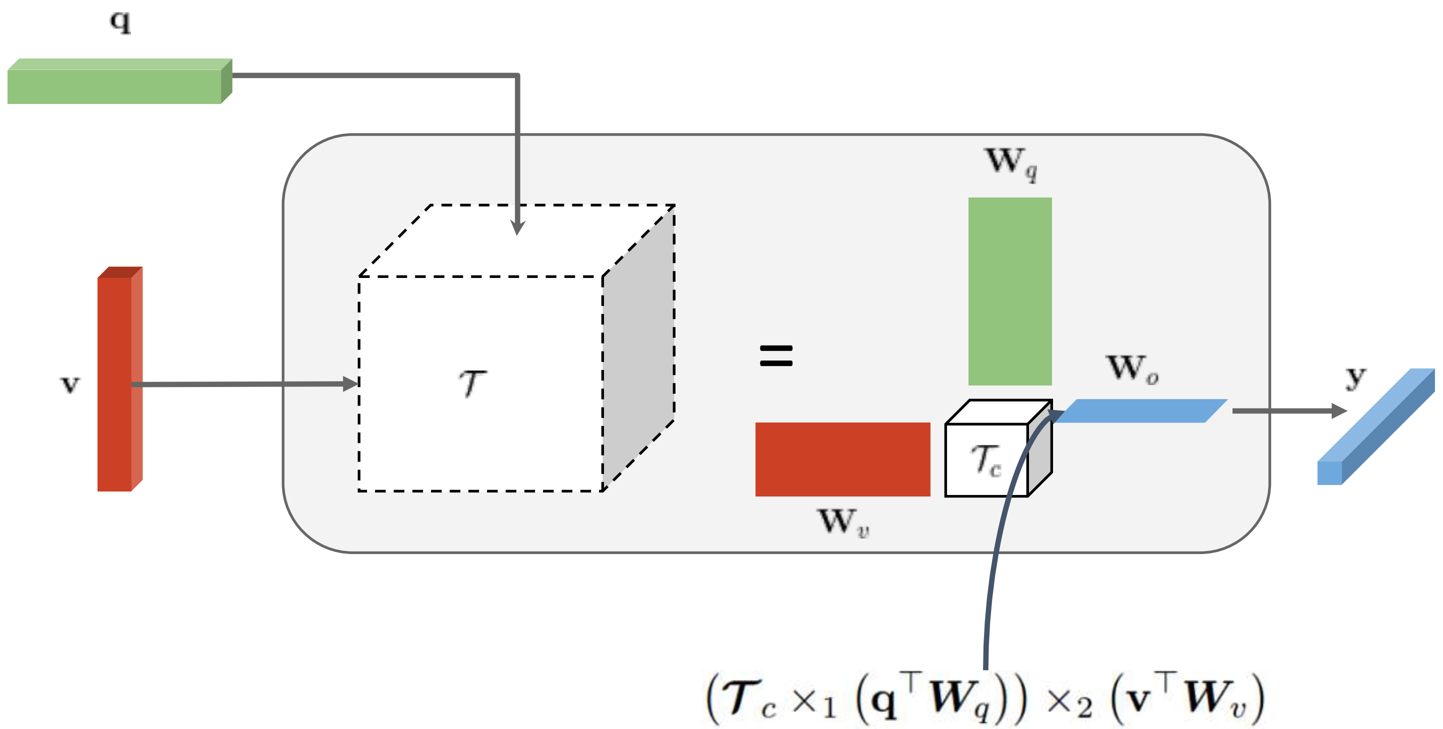
# VQA: bilinear fusion



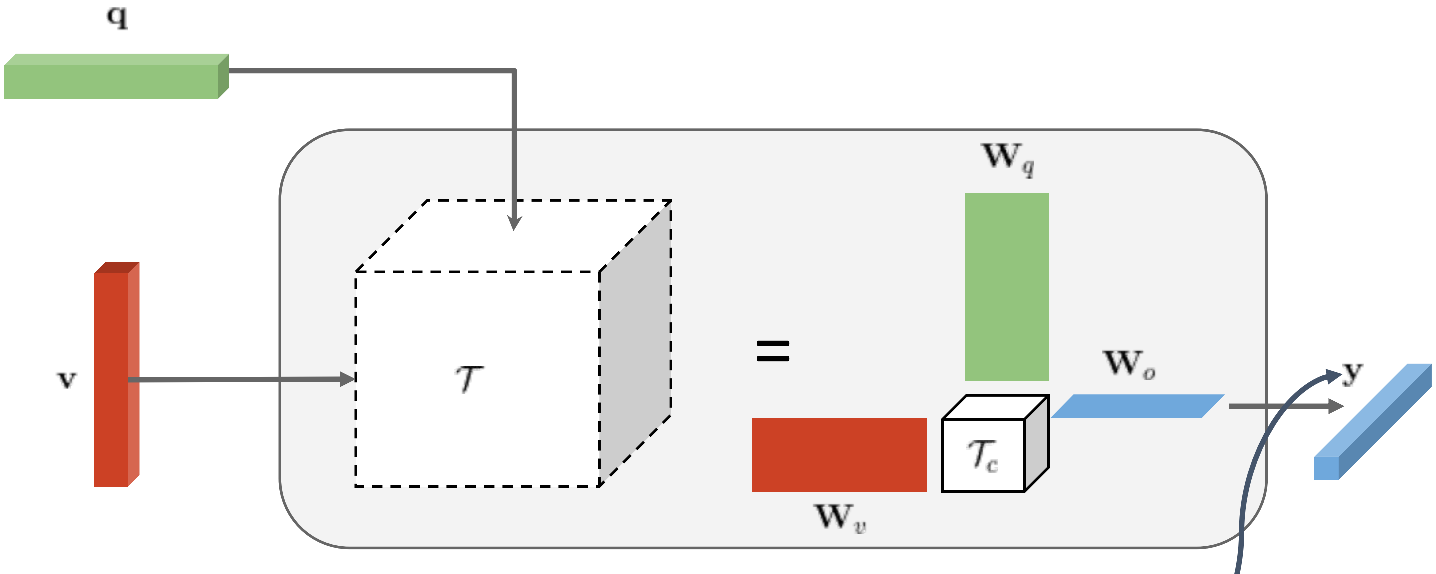
# VQA: bilinear fusion



# VQA: bilinear fusion



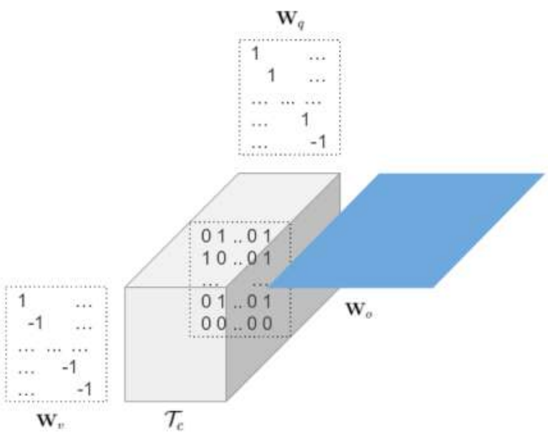
# VQA: bilinear fusion




$$y = \left(\left(\boldsymbol{\mathcal{T}}_c \times_1 \left(\mathbf{q}^\top \boldsymbol{W}_q\right)\right) \times_2 \left(\mathbf{v}^\top \boldsymbol{W}_v\right)\right) \times_3 \boldsymbol{W}_o$$

# VQA: bilinear fusion

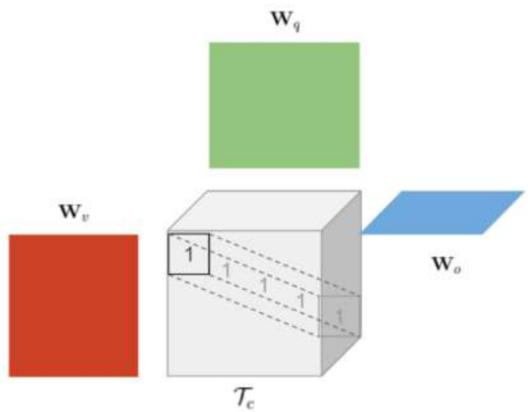
## Other ways of structuring the tensor of parameters




Compact Bilinear

Pooling

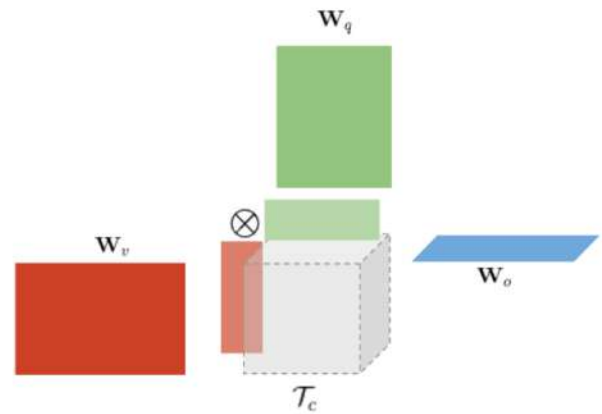
(MCB)




Low-Rank

Bilinear Pooling

(MLB)




Tucker

Decomposition

(MUTAN)

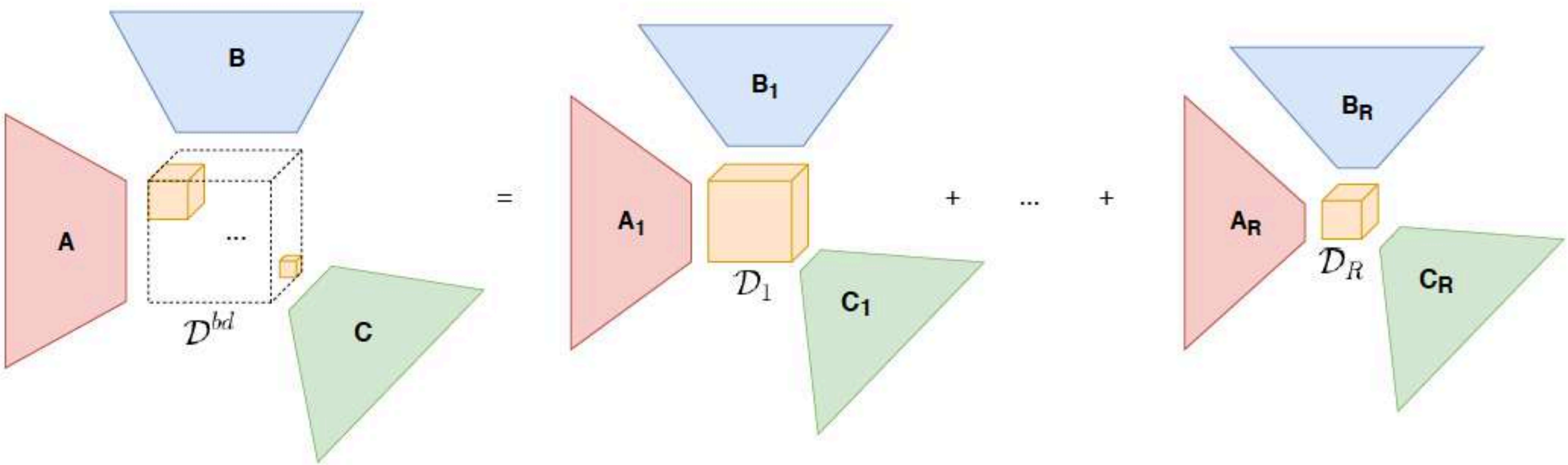
Ben-younes H.* Cadene R.*, Thome N., Cord M., *MUTAN: Multimodal Tucker Fusion for Visual Question Answering,* ICCV 2017

# VQA: bilinear fusion [AAAI 2019]

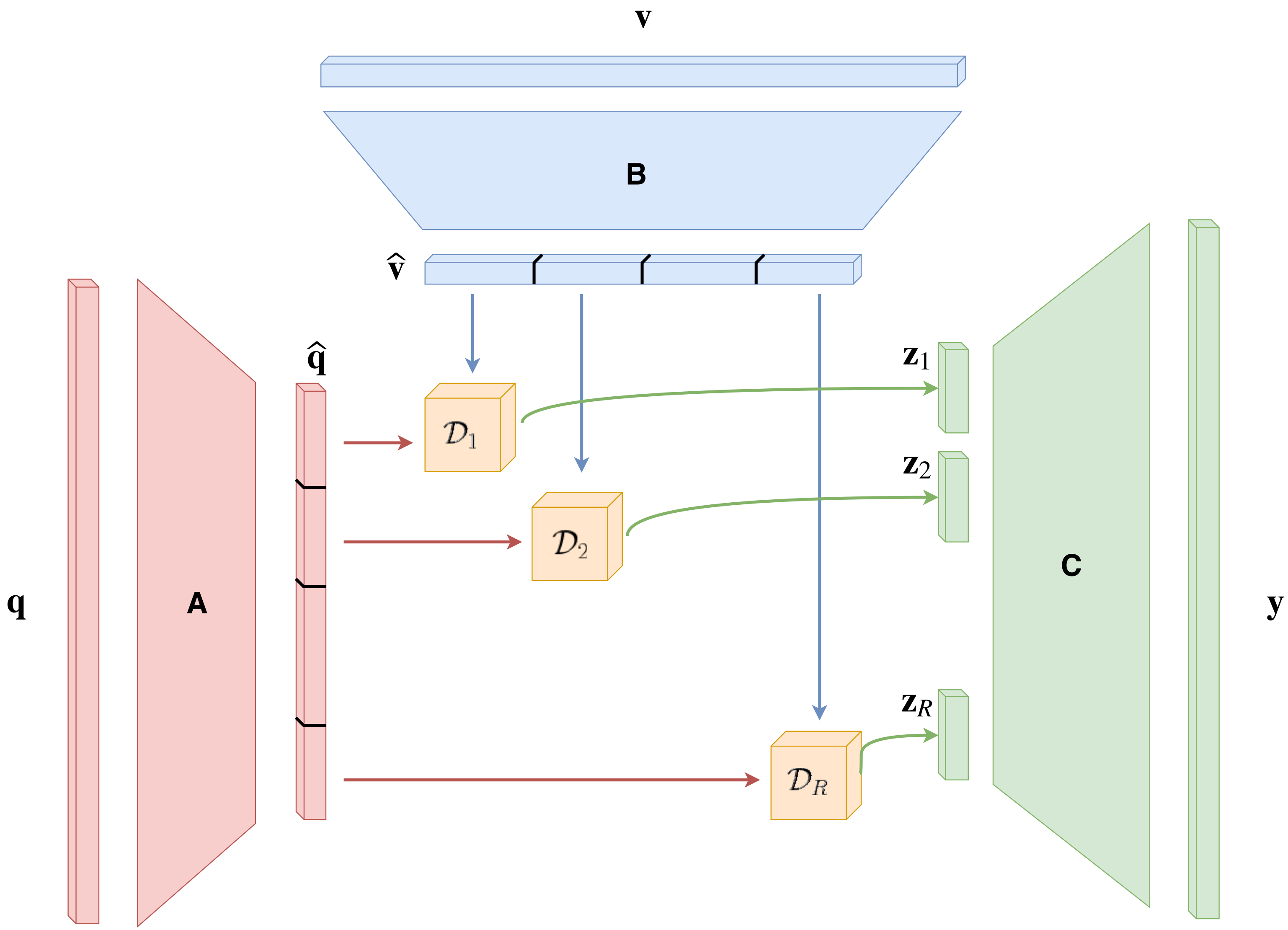
BLOCK [Ben-Younes et al. 2019], extension with a structured $\mathcal{T}$ using a block-term decomposition [De Lathauwer 2008]:

$$\mathcal{T} := \sum_{r=1}^{R} \mathcal{D}_r \times_1 \boldsymbol{A}_r \times_2 \boldsymbol{B}_r \times_3 \boldsymbol{C}_r \qquad (1)$$

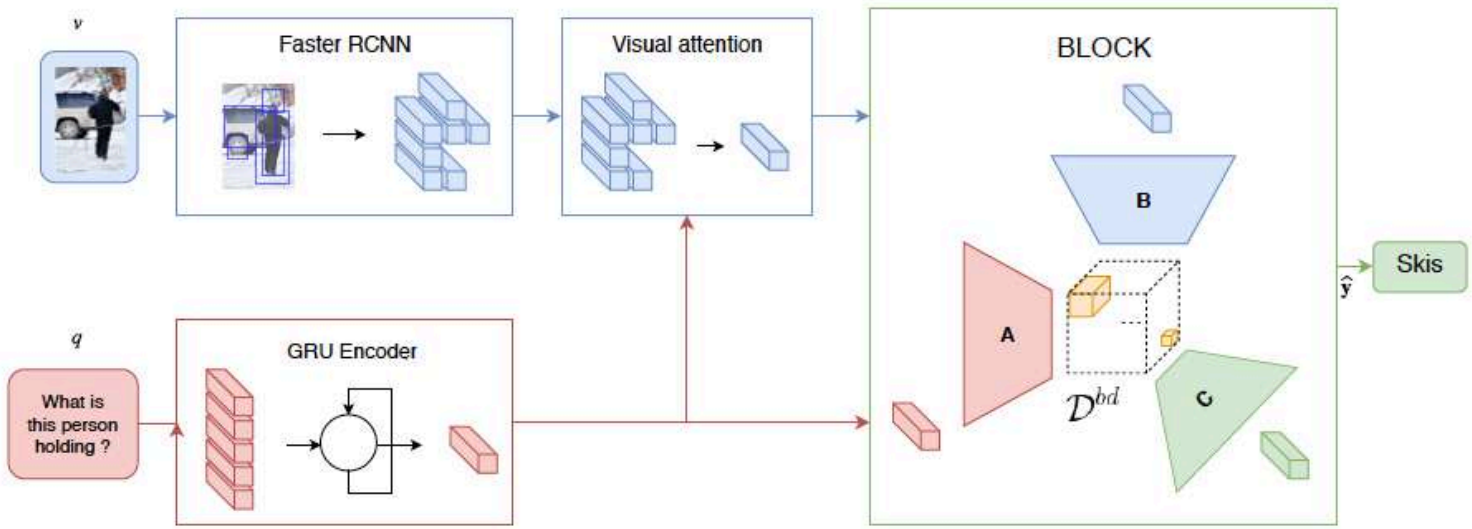
$\mathcal{D}_r \in \mathbb{R}^{L \times M \times N}$, $\boldsymbol{A}_r \in \mathbb{R}^{d_q \times L}$, $\boldsymbol{B}_r \in \mathbb{R}^{d_v \times M}$ and $\boldsymbol{C}_r \in \mathbb{R}^{d_o \times N}$



# VQA: BLOCK fusion [AAAI 2019]



# Classical attention architecture:



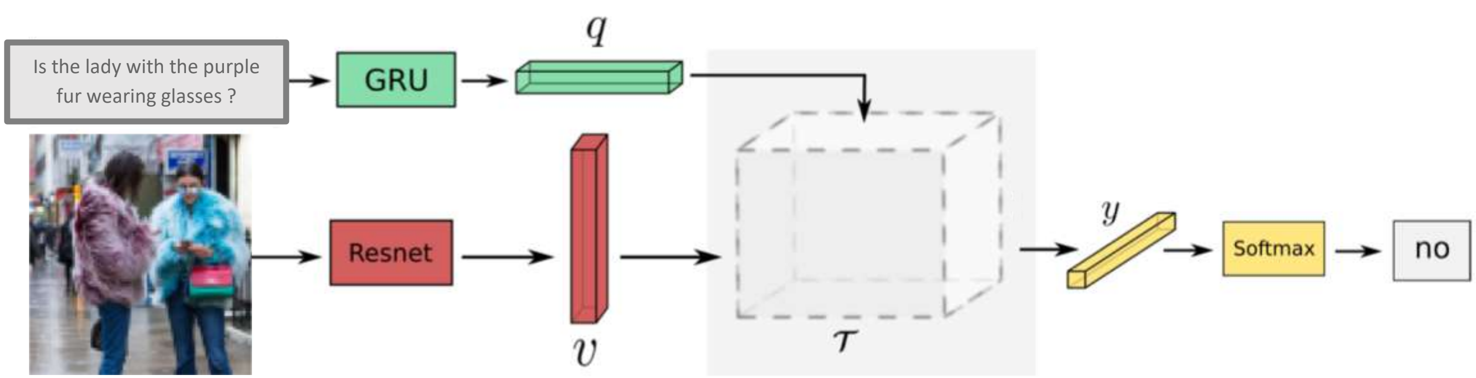

# Comparing fusion schemes on the VQA Dataset 2.0

| | 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** |

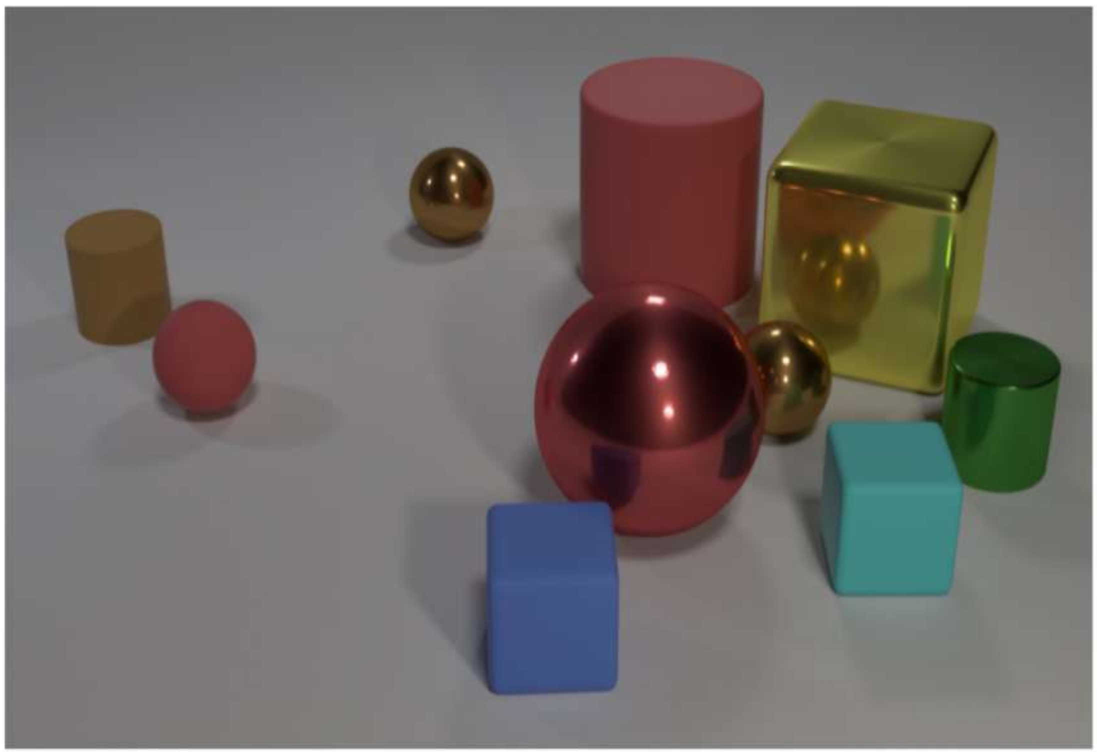
# VQA: bilinear fusion

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



# Reasoning in VQA



**Q:** Are there an equal number of large things and metal spheres?

# VQA: reasoning

What is reasoning (for VQA)?

**Attentional reasoning**

**Relational reasoning**

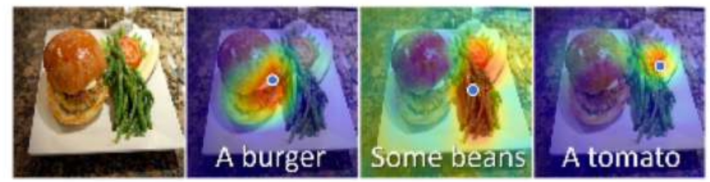
**Iterative reasoning**

**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**

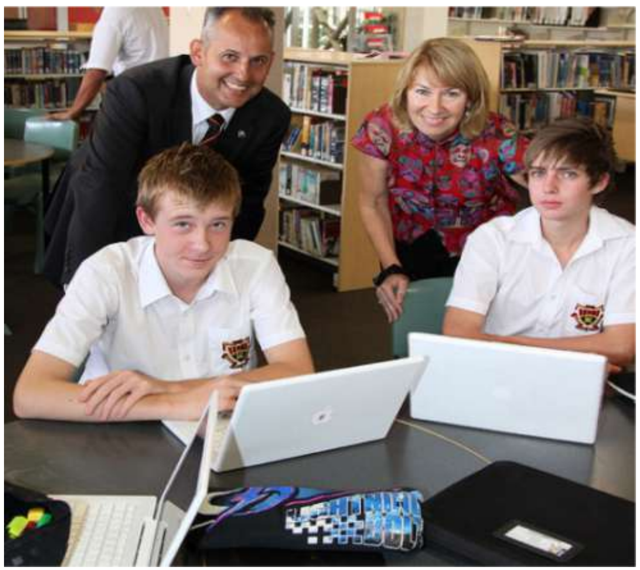
**Compositional reasoning**

# VQA: attentional reasoning

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

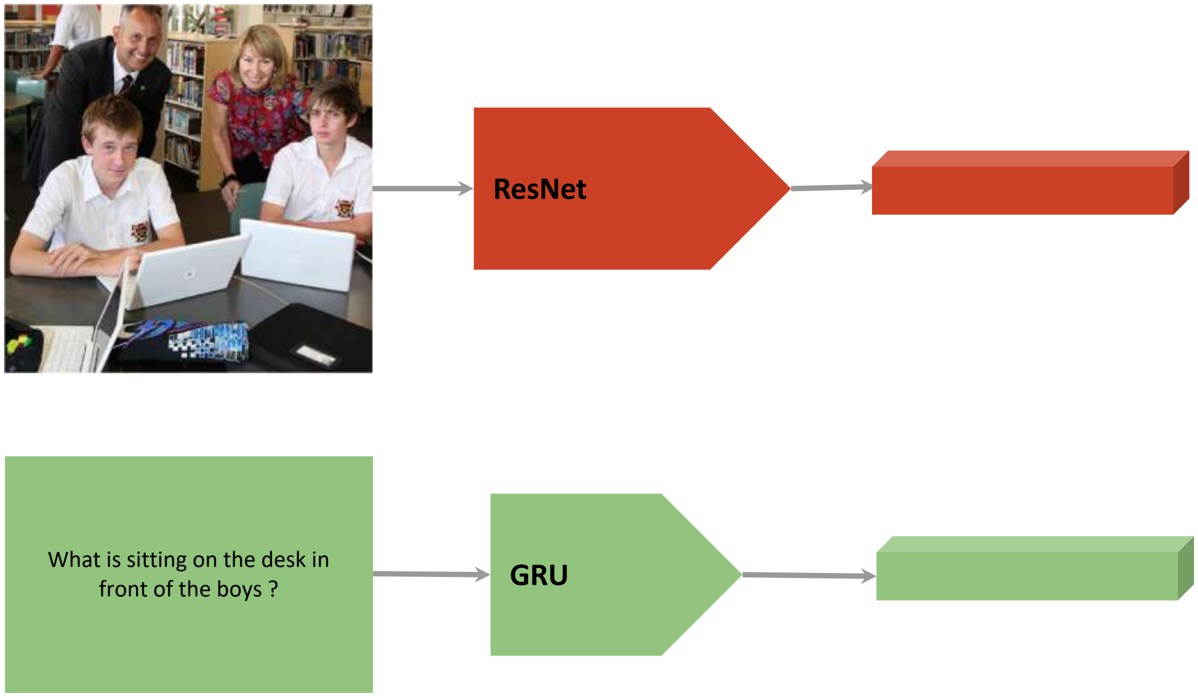
- Each region scored according to the question

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

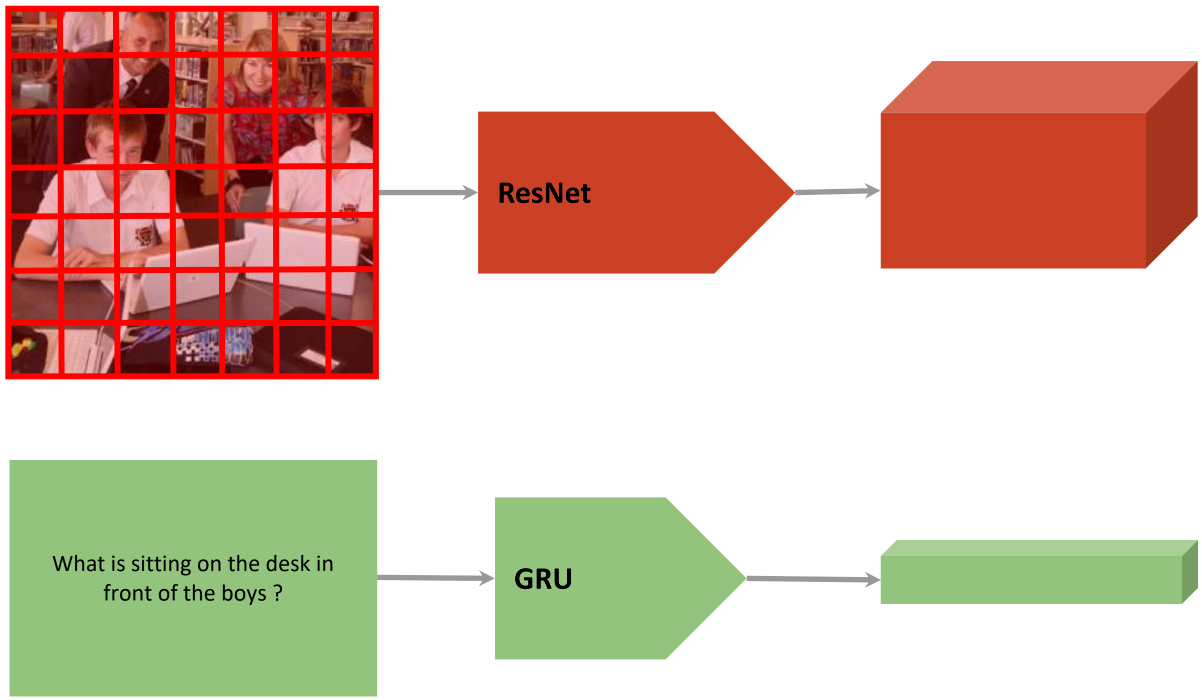


- Representation = sum of all (weighted) embeddings

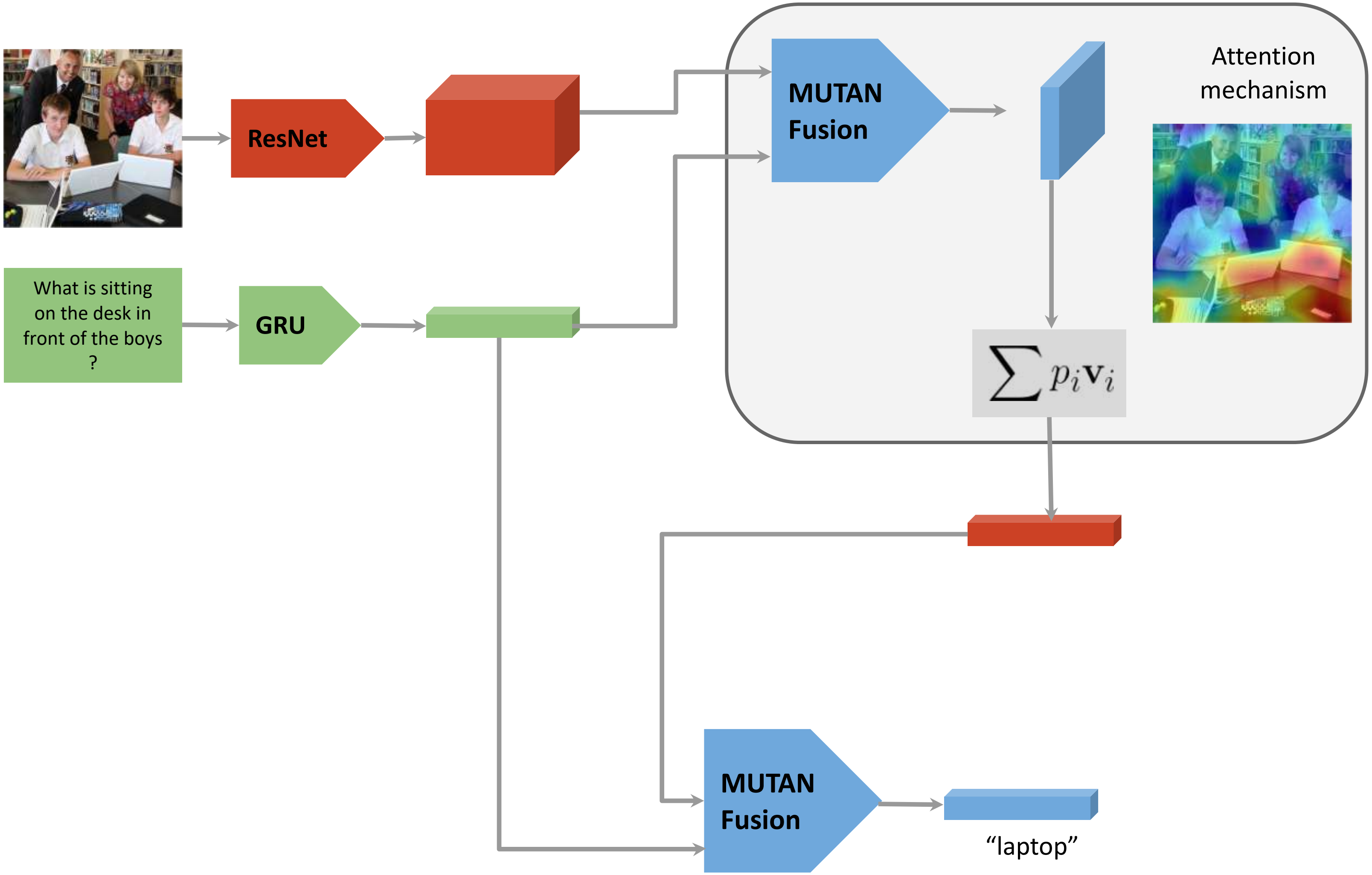
# VQA: attentional reasoning



# VQA: attentional reasoning

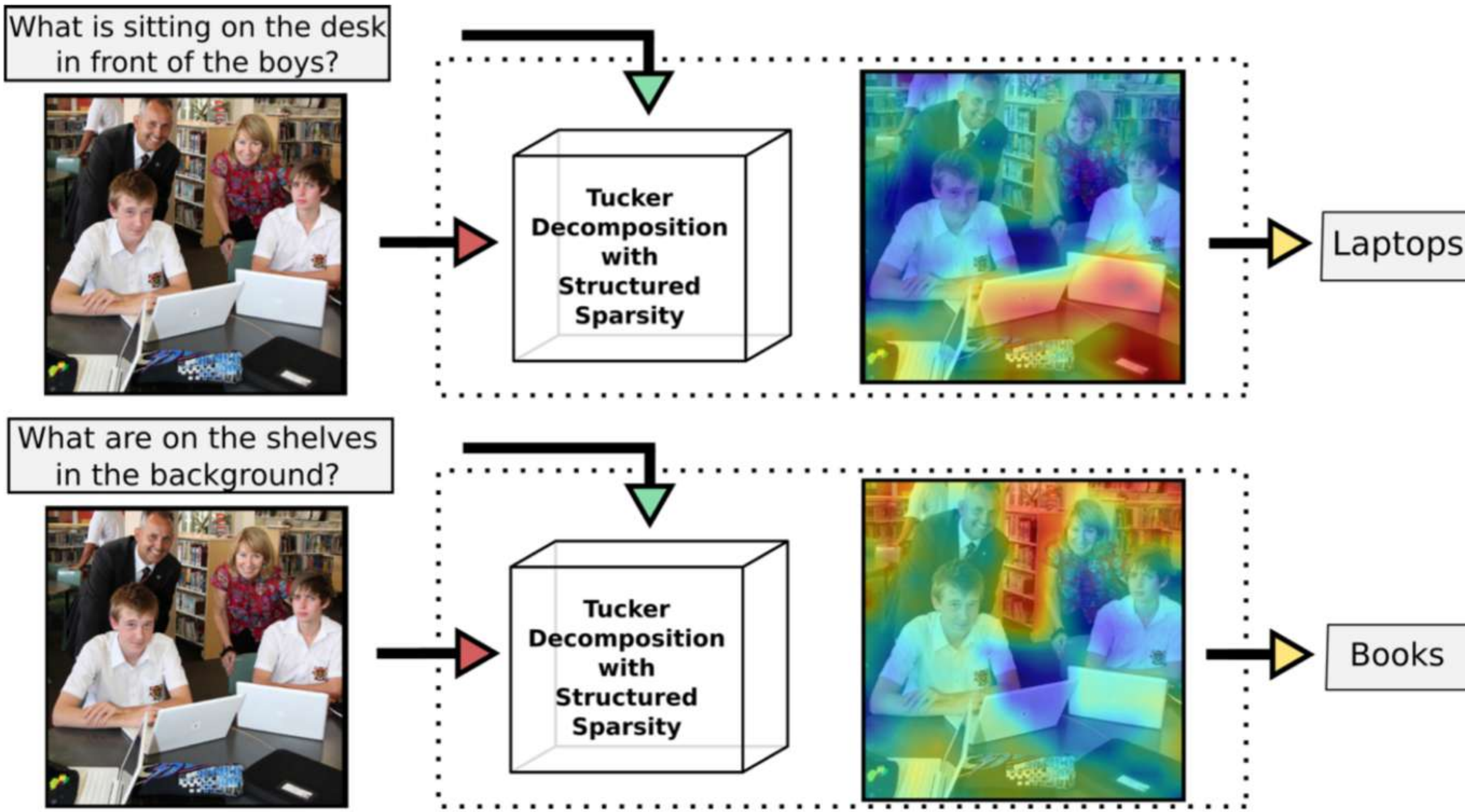


# VQA: attentional reasoning




Attentional glimpse in most of recent strategies [MLB, MCB, *MUTAN*…]

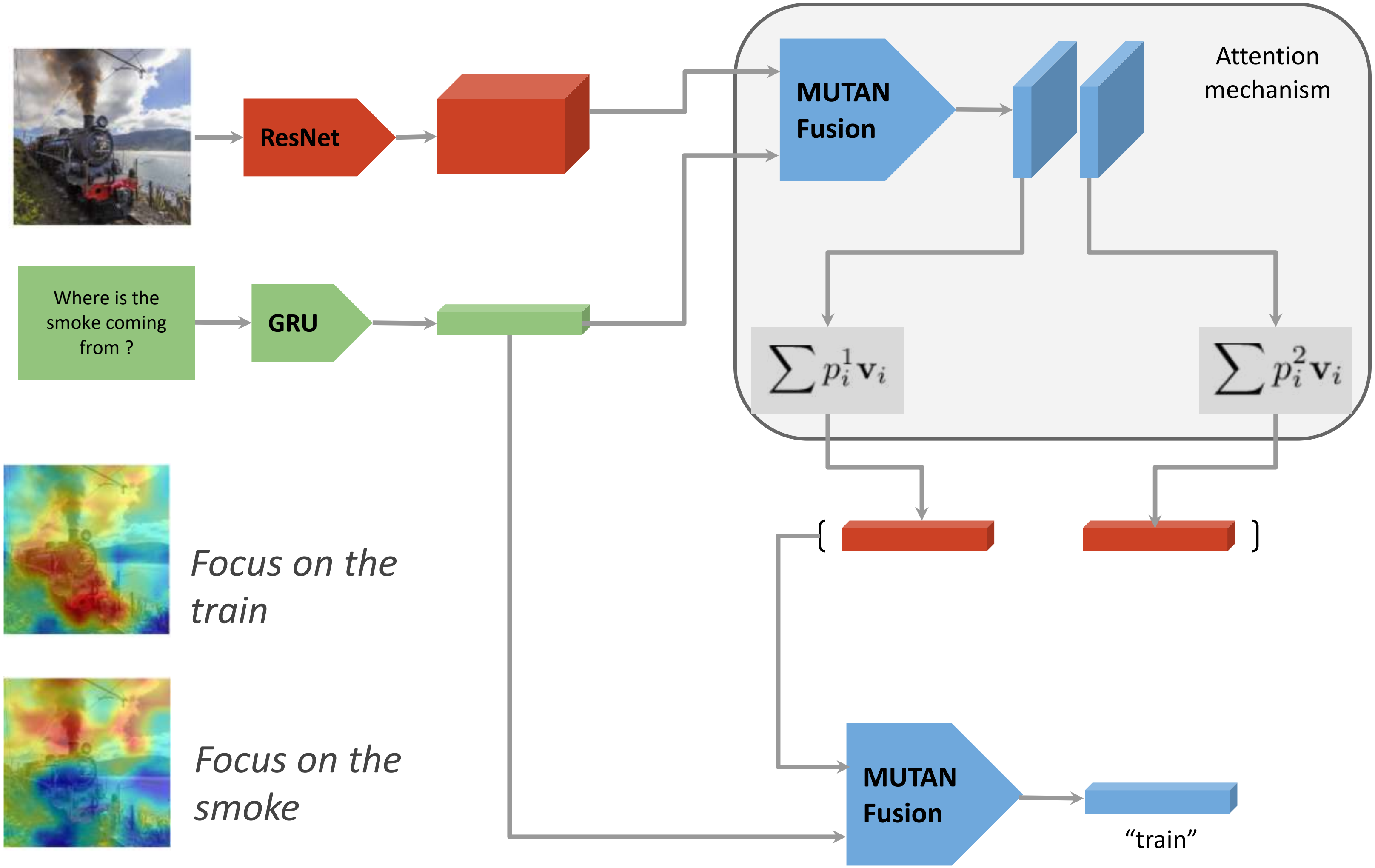
# VQA: attentional reasoning



# VQA: attentional reasoning

## Focusing on multiple regions: Multi-glimpse attention

# VQA: attentional reasoning with Multi-glimpse attention



# 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

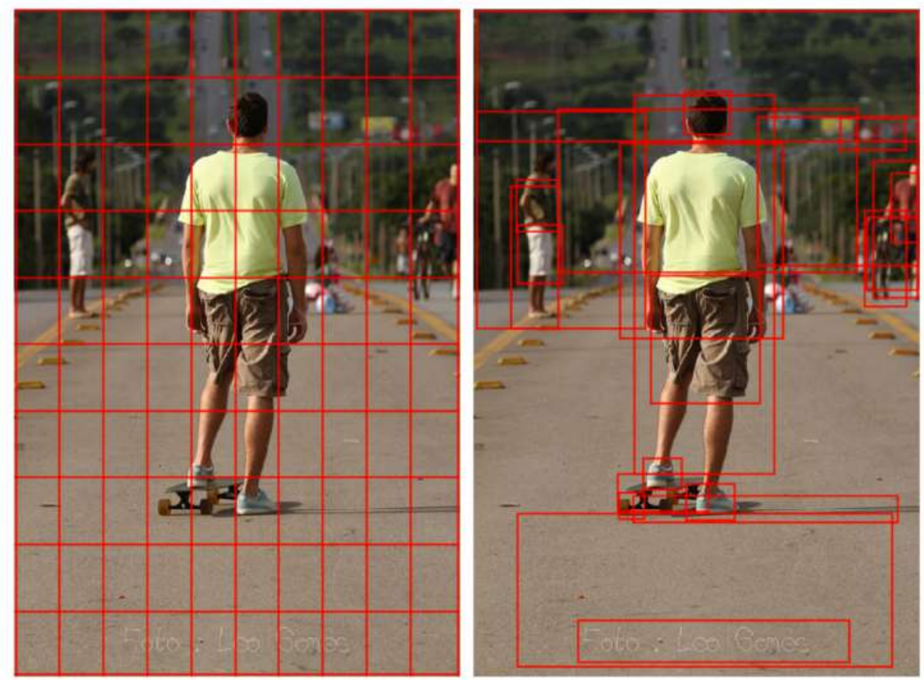
Evaluation on VQA dataset: Best MUTAN score of 67.36% on test-std

Human performances about 83% on this dataset

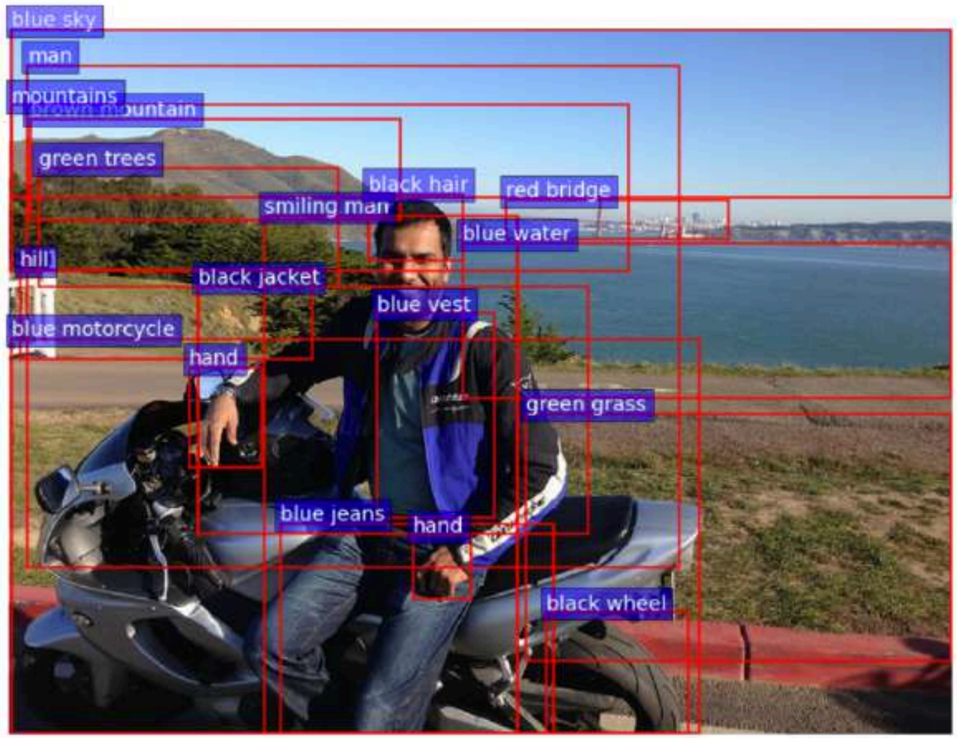
The winner of the VQA Challenge in CVPR 2017 (and CVPR 2018) integrates adaptive grid selection from additional region detection learning process

**Bottom-Up and Top-Down Attention for Image Captioning and VQA**

Peter Anderson[1,*], Xiaodong He[2], Chris Buehler[2], Damien Teney[3]
Mark Johnson[4], Stephen Gould[1], Lei Zhang[2]



# 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)?

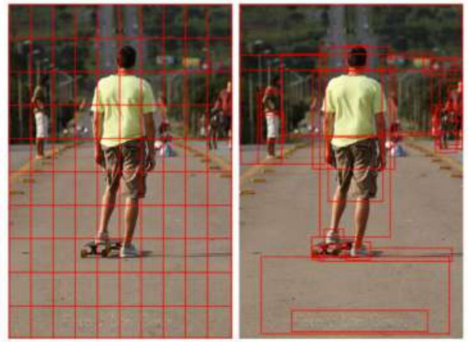
**Attentional reasoning**: given a certain context (i.e. Q), focus only on the relevant subparts of the image

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

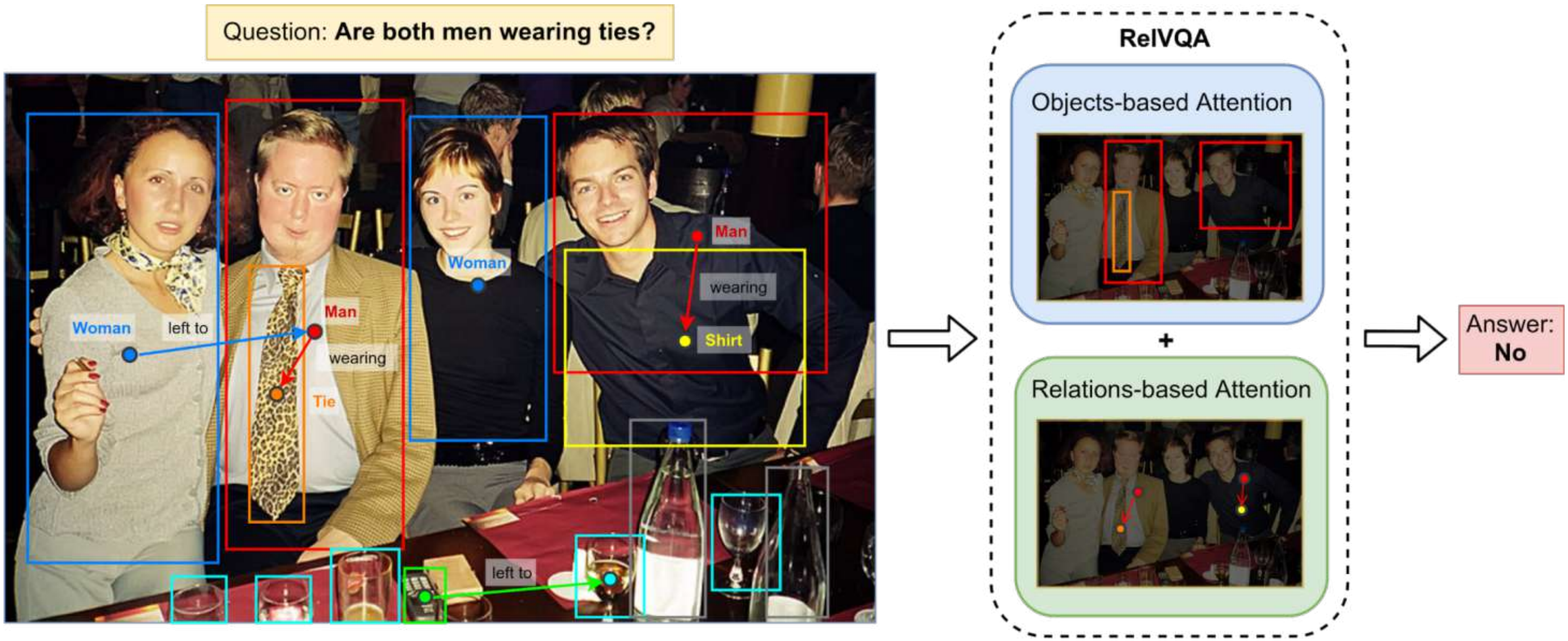
**Iterative reasoning**

**Compositional reasoning**

# Bottom-up and Relational reasoning



Determine the answer using relevant objects and relationships



# 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

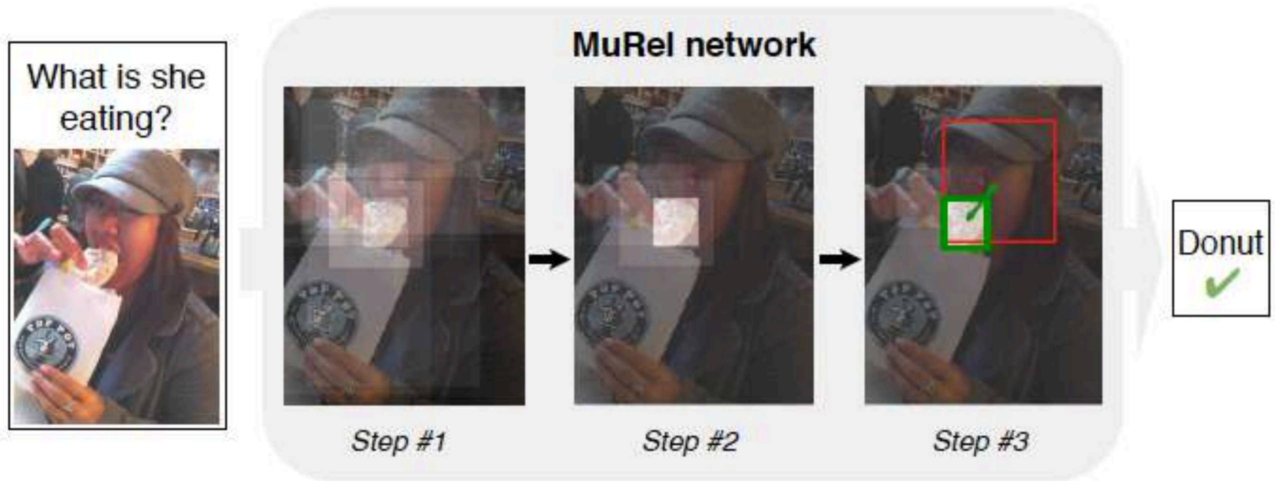
**Stacked attention**: iteratively refining visual attention and question representation

What are sitting in the basket on a bicycle ?

Original Image First Attention Layer Second Attention Layer

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

# Multimodal Relational Reasoning for VQA

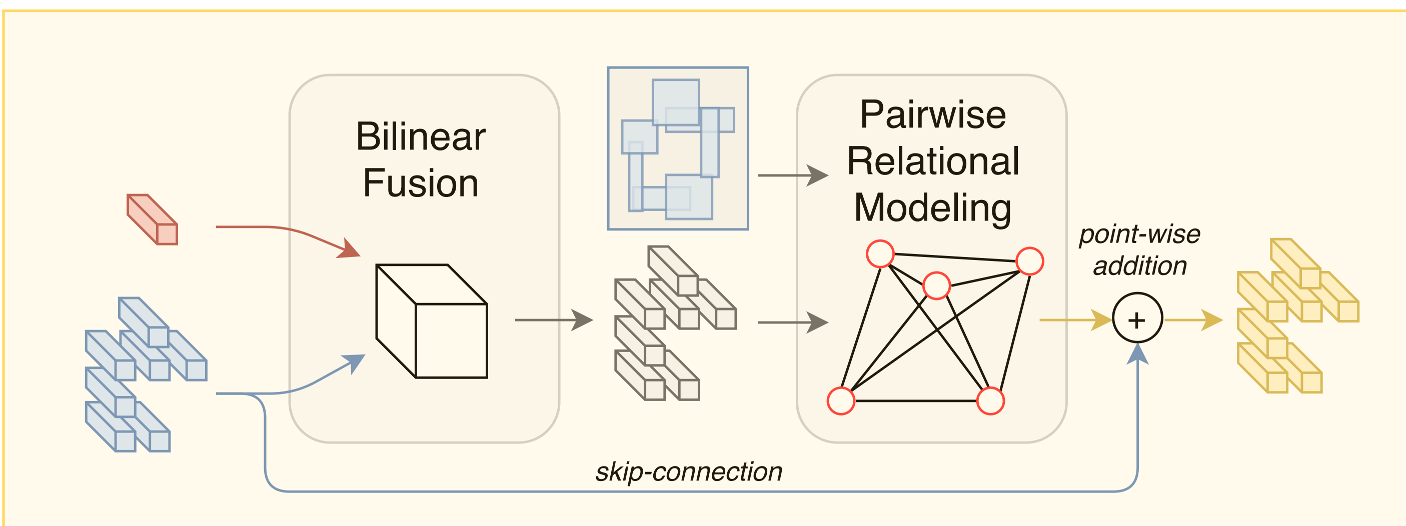



MUREL system:

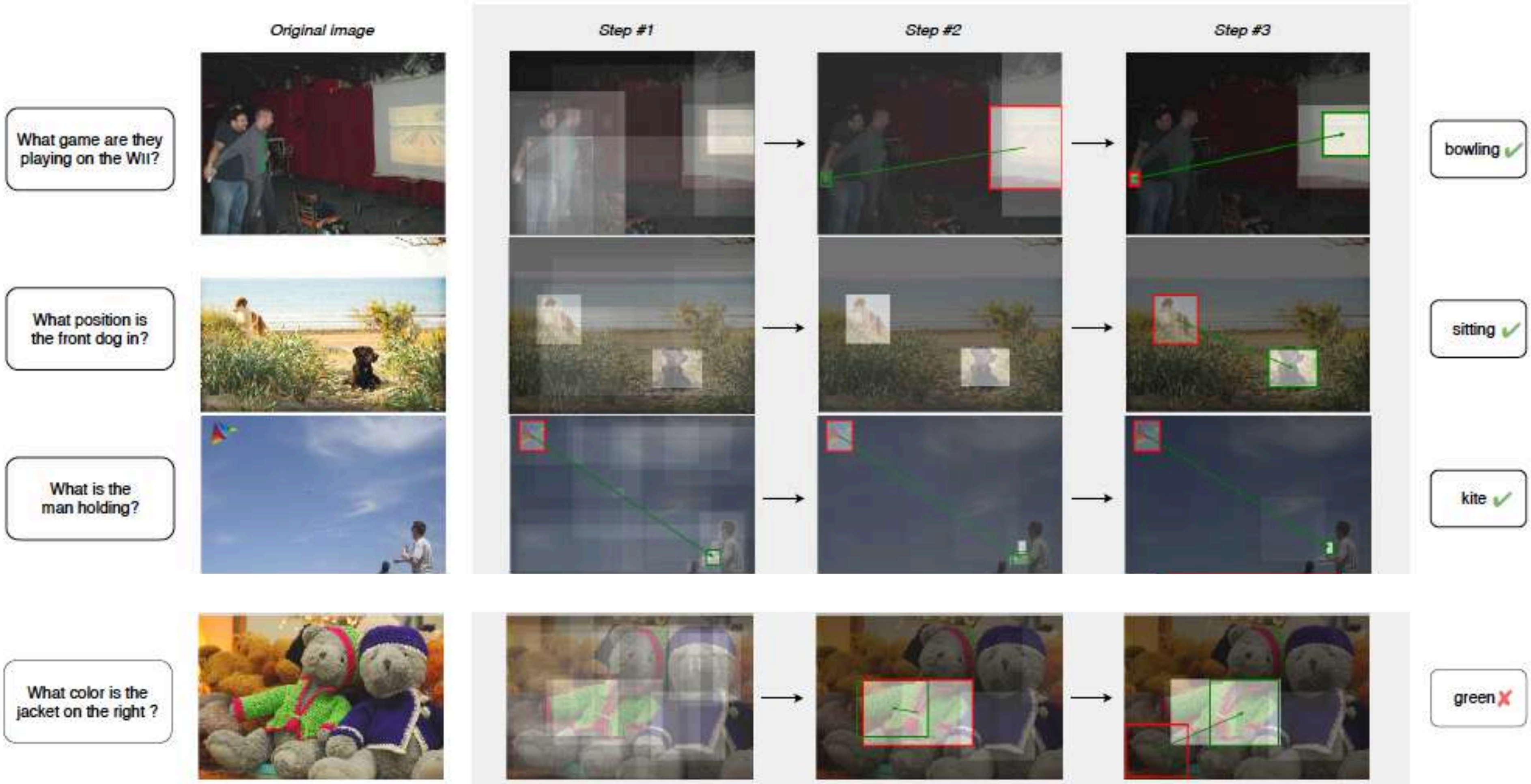
- Vector representation for Attention process
- Spatial and semantic contexts to model relations between image regions
- Iterative process /Multistep reasoning

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

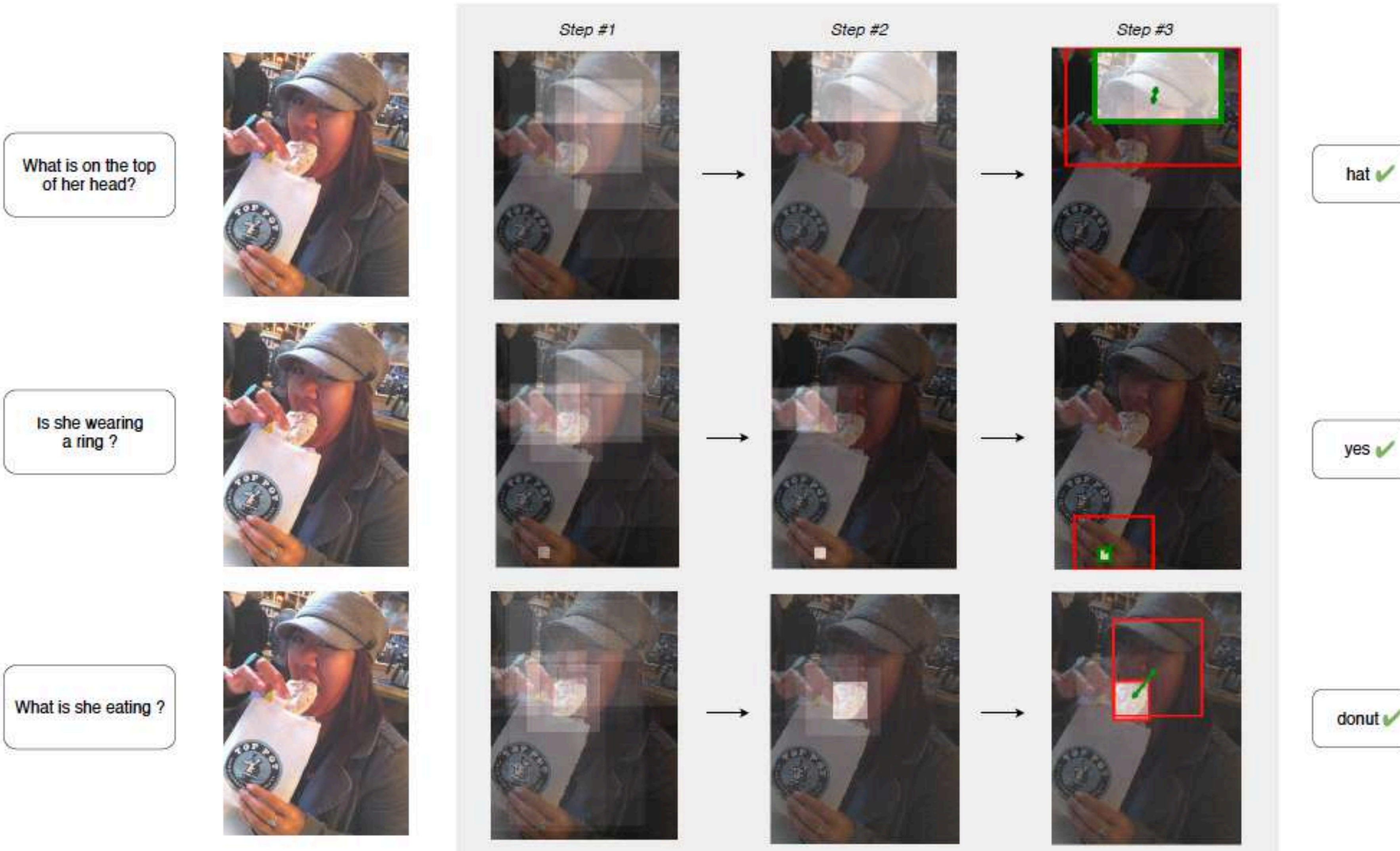
# MuRel: Multimodal Relational Reasoning for VQA

# MuRel: Multimodal Relational Reasoning for VQA



# MuRel: Multimodal Relational Reasoning for VQA



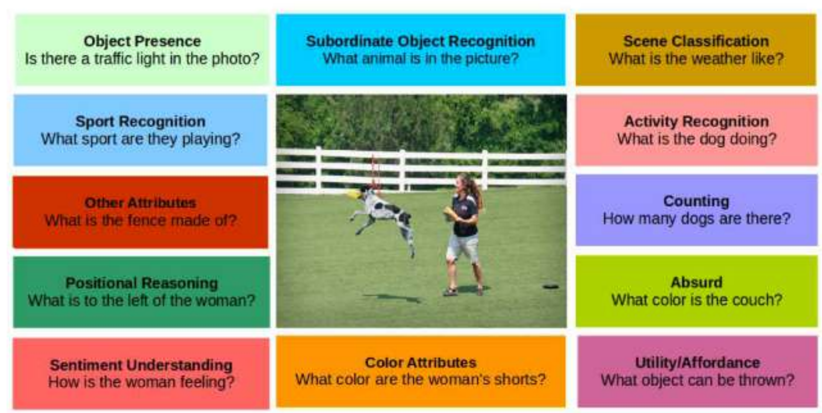
## VQA v2.0 dataset

| Model | test-dev Yes/No | test-dev Num. | test-dev Other | test-dev All | test-std All |
|---|---|---|---|---|---|
| Bottom-up [3] | 81.82 | 44.21 | 56.05 | 65.32 | 65.67 |
| Graph Att. [31] | - | - | - | - | 66.18 |
| MUTAN† [8] | 82.88 | 44.54 | 56.50 | 66.01 | 66.38 |
| MLB† [20] | 83.58 | 44.92 | 56.34 | 66.27 | 66.62 |
| DA-NTN [6] | 84.29 | 47.14 | 57.92 | 67.56 | 67.94 |
| Pythia [40] | - | - | - | 68.05 | - |
| Counter [41] | 83.14 | **51.62** | **58.97** | **68.09** | **68.41** |
| MuRel | **84.77** | 49.84 | 57.85 | 68.03 | **68.41** |


Figure 1: Examples from our balanced VQA dataset.

## TDIUC dataset (12 different categories)

| | RAU* [30] | MCB* [11] | QTA [35] | MuRel |
|---|---|---|---|---|
| Bottom-up | ✗ | ✗ | ✓ | ✓ |
| Scene Reco. | 93.96 | 93.06 | 93.80 | **96.11** |
| Sport Reco. | 93.47 | 92.77 | 95.55 | **96.20** |
| Color Attr. | 66.86 | 68.54 | 60.16 | **74.43** |
| Other Attr. | 56.49 | 56.72 | 54.36 | **58.19** |
| Activity Reco. | 51.60 | 52.35 | 60.10 | **63.83** |
| Pos. Reasoning | 35.26 | 35.40 | 34.71 | **41.19** |
| Object Reco. | 86.11 | 85.54 | 86.98 | **89.41** |
| Absurd | 96.08 | 84.82 | **100.00** | 99.8 |
| Util. and Afford. | 31.58 | **35.09** | 31.48 | 21.43 |
| Object Presence | 94.38 | 93.64 | 94.55 | **95.75** |
| Counting | 48.43 | 51.01 | 53.25 | **61.78** |
| Sentiment | 60.09 | **66.25** | 64.38 | 60.65 |
| Overall (A-MPT) | 67.81 | 67.90 | 69.11 | **71.56** |
| Overall (H-MPT) | 59.00 | **60.47** | 60.08 | 59.30 |
| Overall Accuracy | 84.26 | 81.86 | 85.03 | **88.20** |



# 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


Figure 1: Examples from our balanced VQA dataset.

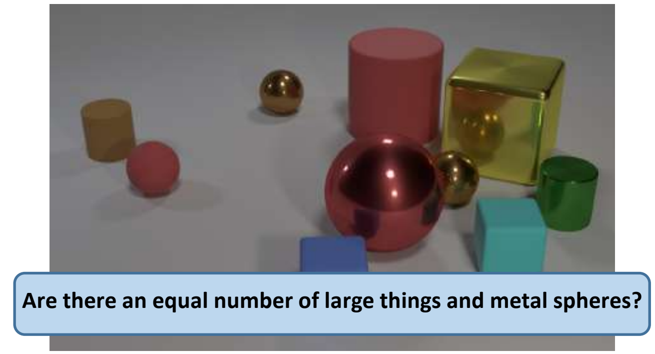




Figure 1: Examples from the new GQA dataset for visual reasoning and compositional question answering:
*Is the **bowl** to the right of the **green apple**?*
*What type of **fruit** in the image is **round**?*
*What color is the **fruit** on the right side, red or **green**?*
*Is there any **milk** in the **bowl** to the left of the **apple**?*

MLIA/Chordettes team:
Matthieu Cord [http://webia.lip6.fr/~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

AAAI 2019 **BLOCK: Bilinear Superdiagonal Fusion for Visual Question Answering and Visual Relationship Detection ,** H. Ben-younes, R. Cadene, 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

Our Deep Recipe Reco on your mobile: [visiir.lip6.fr](http://visiir.lip6.fr)