FRANCE IS AI conference, Second edition

Come for 2 days of roundtables, talks, and startup competition with Researchers, Entrepreneurs, Executives and VCs

Oct 5-6 2017, Wework Paris

Visual Question Answering: a new Vision and Language task

Matthieu Cord

UPMC (Sorbonne Univ.)/LIP6/MLIA - UMR CNRS

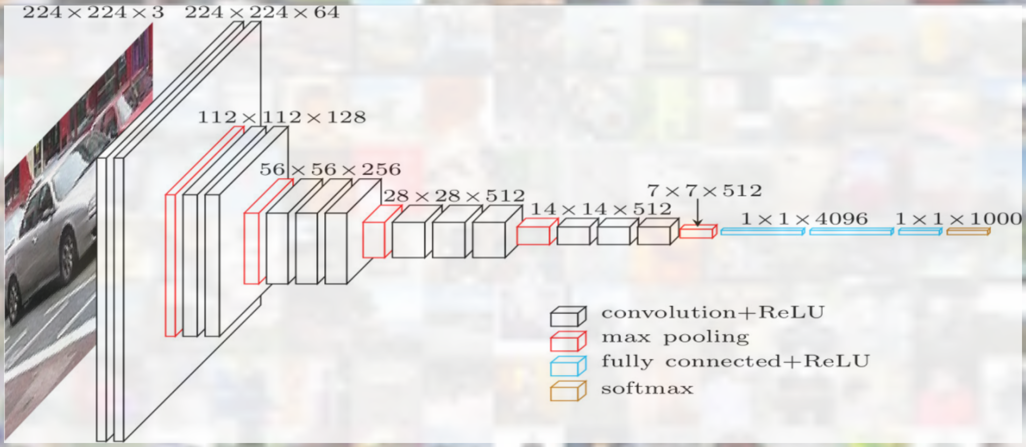
UPMC SORBONNE UNIVERSITÉS

MLIA

Visual Question Answering

MLIA team:

- About 35 researchers and PhD students (head P. Gallinari) on Machine Learning/ Deep Learning
- Computer Vision side:
    - VQA: MUTAN paper at ICCV17 PhD Hedi Benyounes (with HEURITECH) and Rémi Cadène (Labex SMART)



Visual Question Answering

Question Answering:

Is Paul in the room?

Visual Question Answering

Visual Question Answering:

Is Paul in the room?

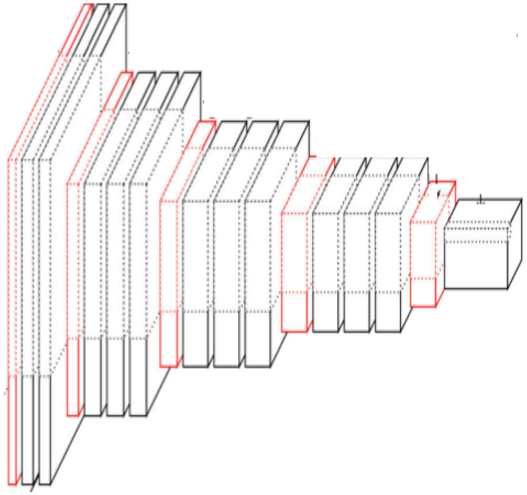
+

Visual Question Answering

Visual Question Answering:

Is Paul in the room?

+



Yes/No

On the left

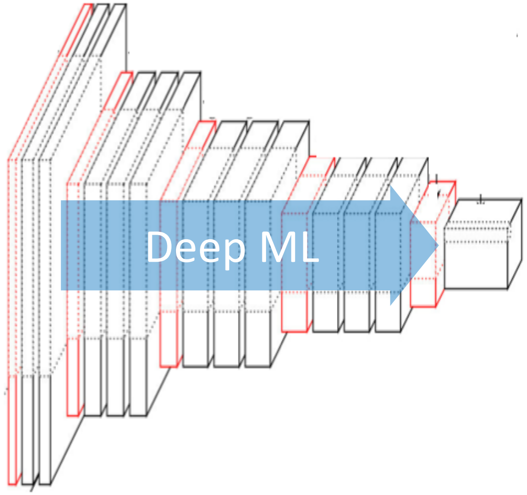
At the back …

Visual Question Answering

Visual Question Answering:

Is Paul in the room?

+



Yes/No

On the left

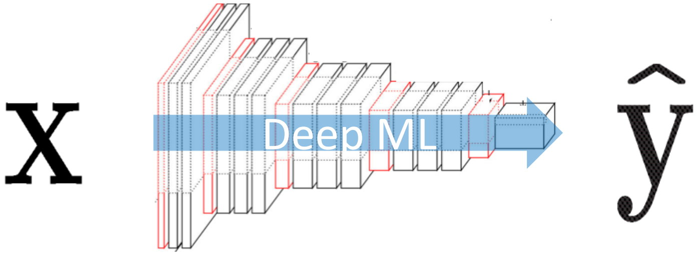
At the back …

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

Context: Vision and Language

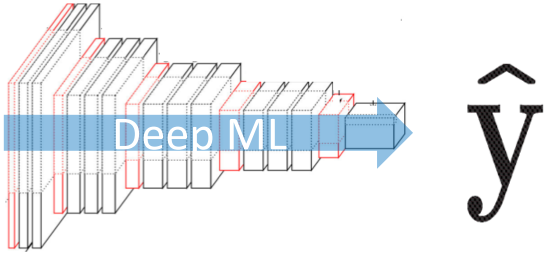
Classification: from Image to keywords/labels



Available Web demo (@Clarifai)

Context: Vision and Language

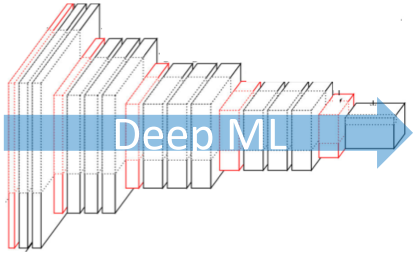
Classification: from Image to keywords/labels



Available Web demo (@Clarifai)

Context: Vision and Language

Classification: from Image to keywords/labels

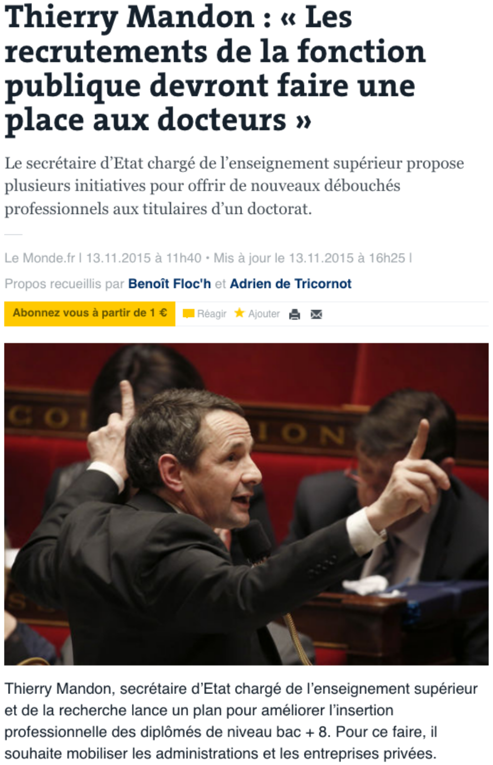


Restaurant
People
Table
Inside
…
Results> 95%

Available Web demo (@Clarifai)

Context: Vision and Language

Classification: from Image to keywords/labels



Thierry Mandon : « Les recrutements de la fonction publique devront faire une place aux docteurs »

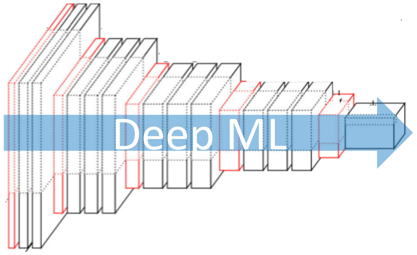
Le secrétaire d'Etat chargé de l'enseignement supérieur propose plusieurs initiatives pour offrir de nouveaux débouchés professionnels aux titulaires d'un doctorat.

Le Monde.fr | 13.11.2015 à 11h40 • Mis à jour le 13.11.2015 à 16h25 |

Propos recueillis par **Benoît Floc'h** et **Adrien de Tricornot**

Abonnez vous à partir de 1 € Réagir ★ Ajouter

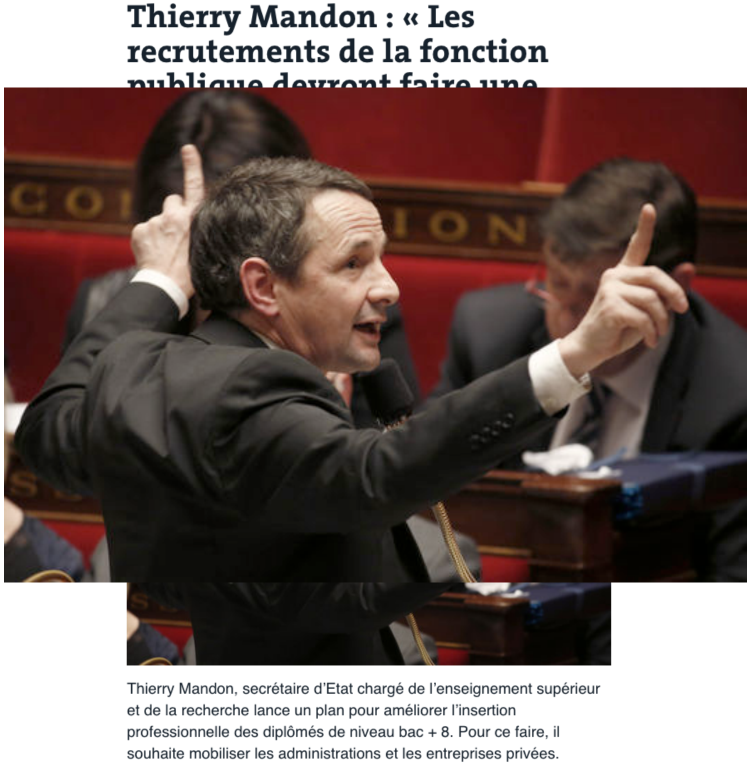
Thierry Mandon, secrétaire d'Etat chargé de l'enseignement supérieur et de la recherche lance un plan pour améliorer l'insertion professionnelle des diplômés de niveau bac + 8. Pour ce faire, il souhaite mobiliser les administrations et les entreprises privées.



Available Web demo (@Clarifai)

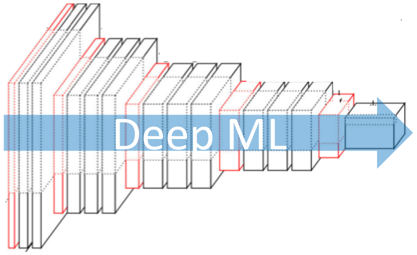
Context: Vision and Language

Classification: from Image to keywords/labels



Thierry Mandon : « Les recrutements de la fonction publique devront faire une

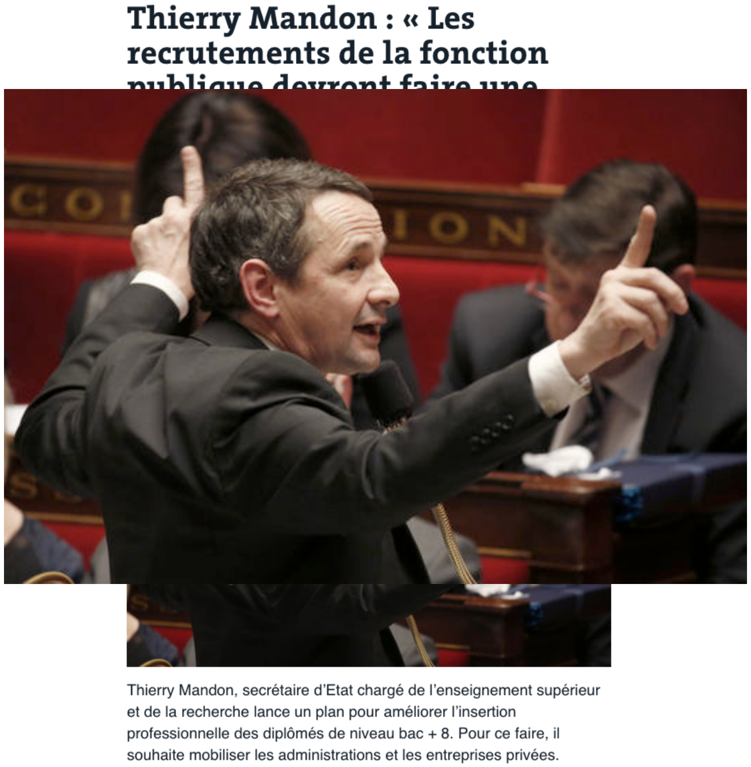
Thierry Mandon, secrétaire d'Etat chargé de l'enseignement supérieur et de la recherche lance un plan pour améliorer l'insertion professionnelle des diplômés de niveau bac + 8. Pour ce faire, il souhaite mobiliser les administrations et les entreprises privées.



Available Web demo (@Clarifai)

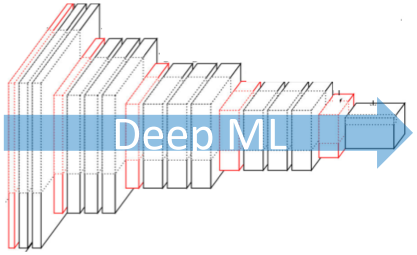
Context: Vision and Language

Classification: from Image to keywords/labels



Thierry Mandon : « Les recrutements de la fonction publique devront faire une

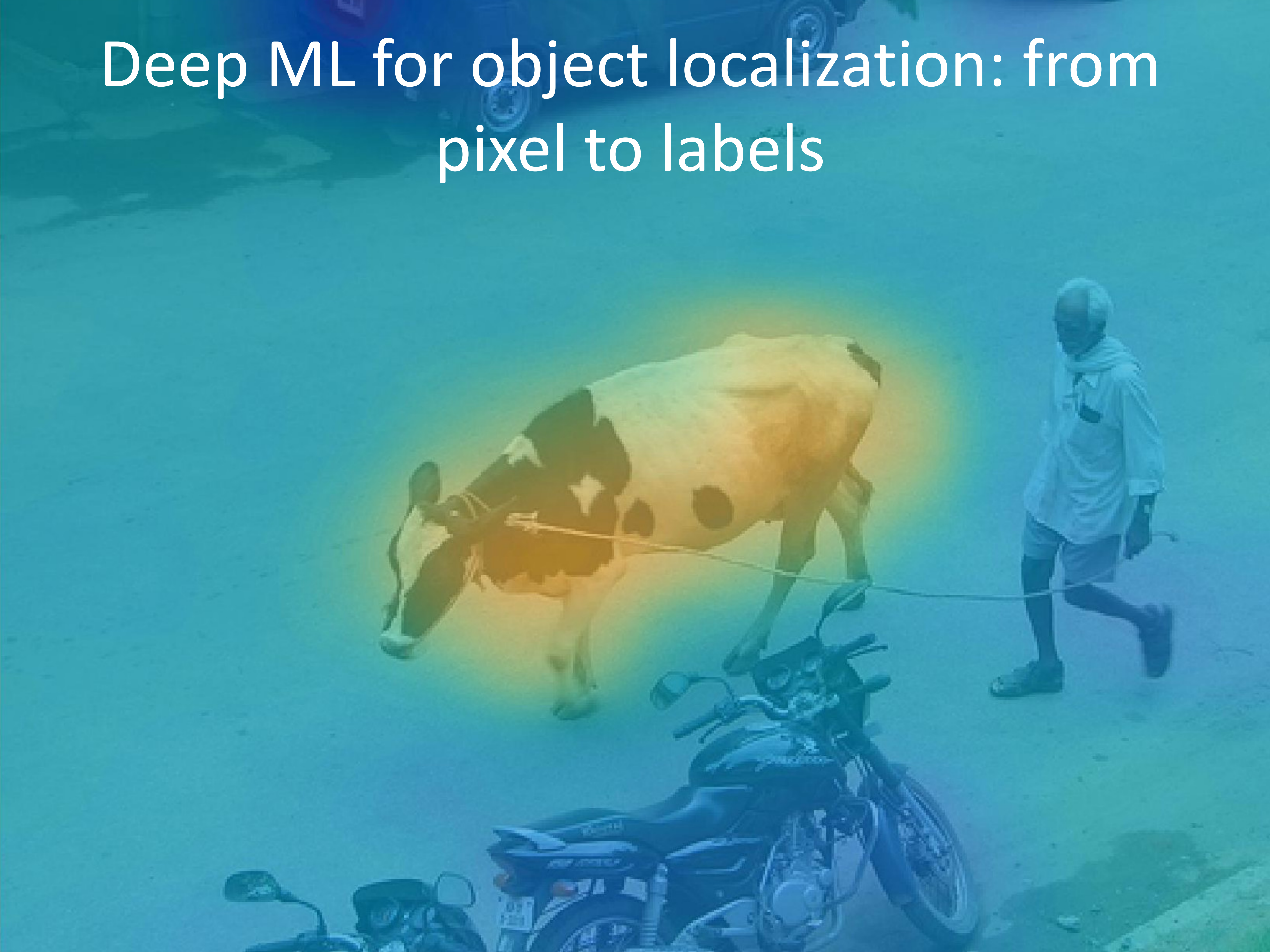
Thierry Mandon, secrétaire d'Etat chargé de l'enseignement supérieur et de la recherche lance un plan pour améliorer l'insertion professionnelle des diplômés de niveau bac + 8. Pour ce faire, il souhaite mobiliser les administrations et les entreprises privées.



Leader
Administration
Election
People
Chair
Results> 95%

Available Web demo (@Clarifai)

Deep ML for object localization: from pixel to labels

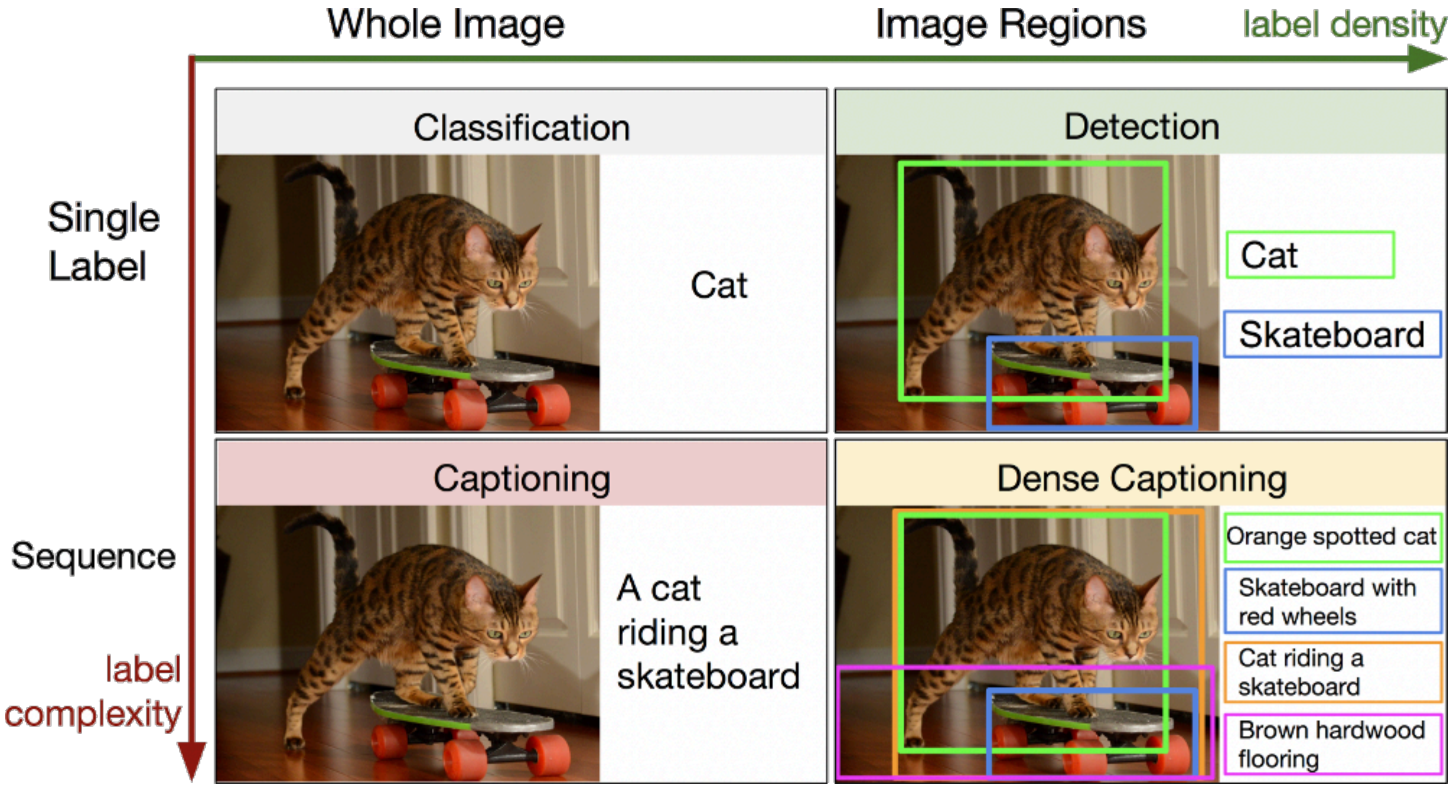


Deep ML for object localization: from pixel to labels

Deep ML for object localization: from pixel to labels

Deep ML for object localization: from pixel to labels

Context: Vision and Language



Language description/complexity

@Feifei

Vision and Language: from keywords to sentence …

Context: Vision and Language

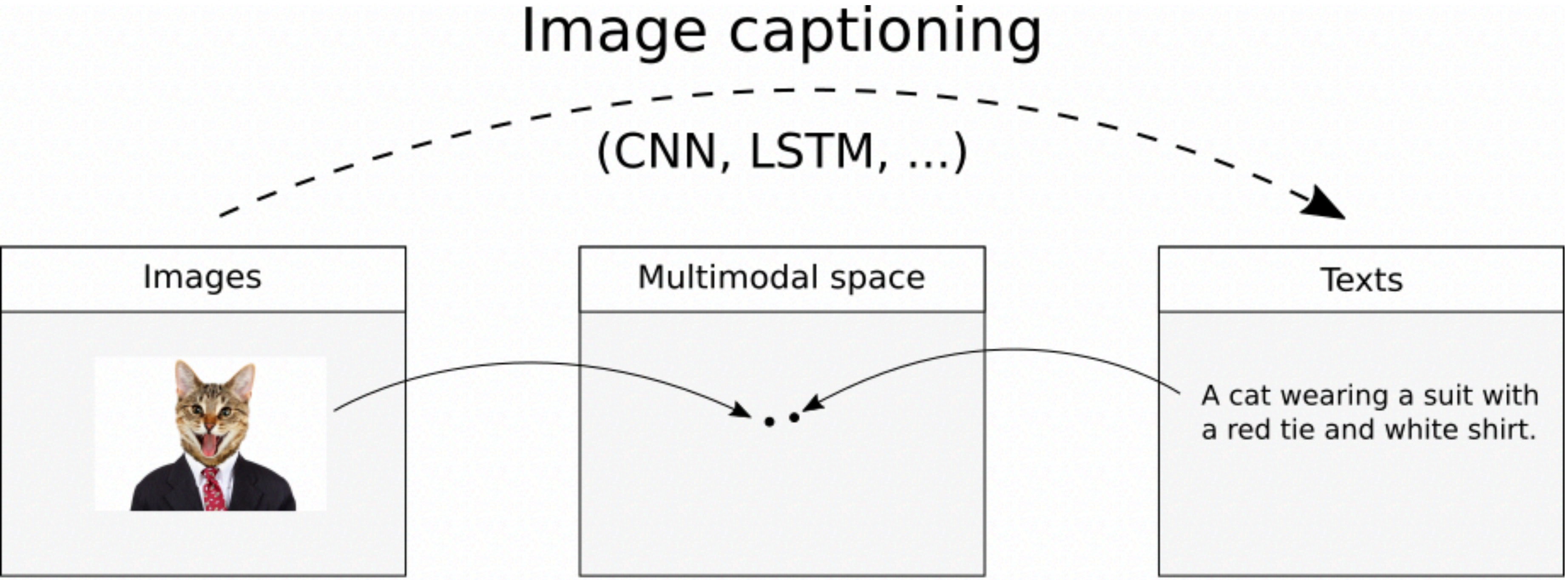


Image captioning

(CNN, LSTM, ...)

Images

Multimodal space

Texts

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

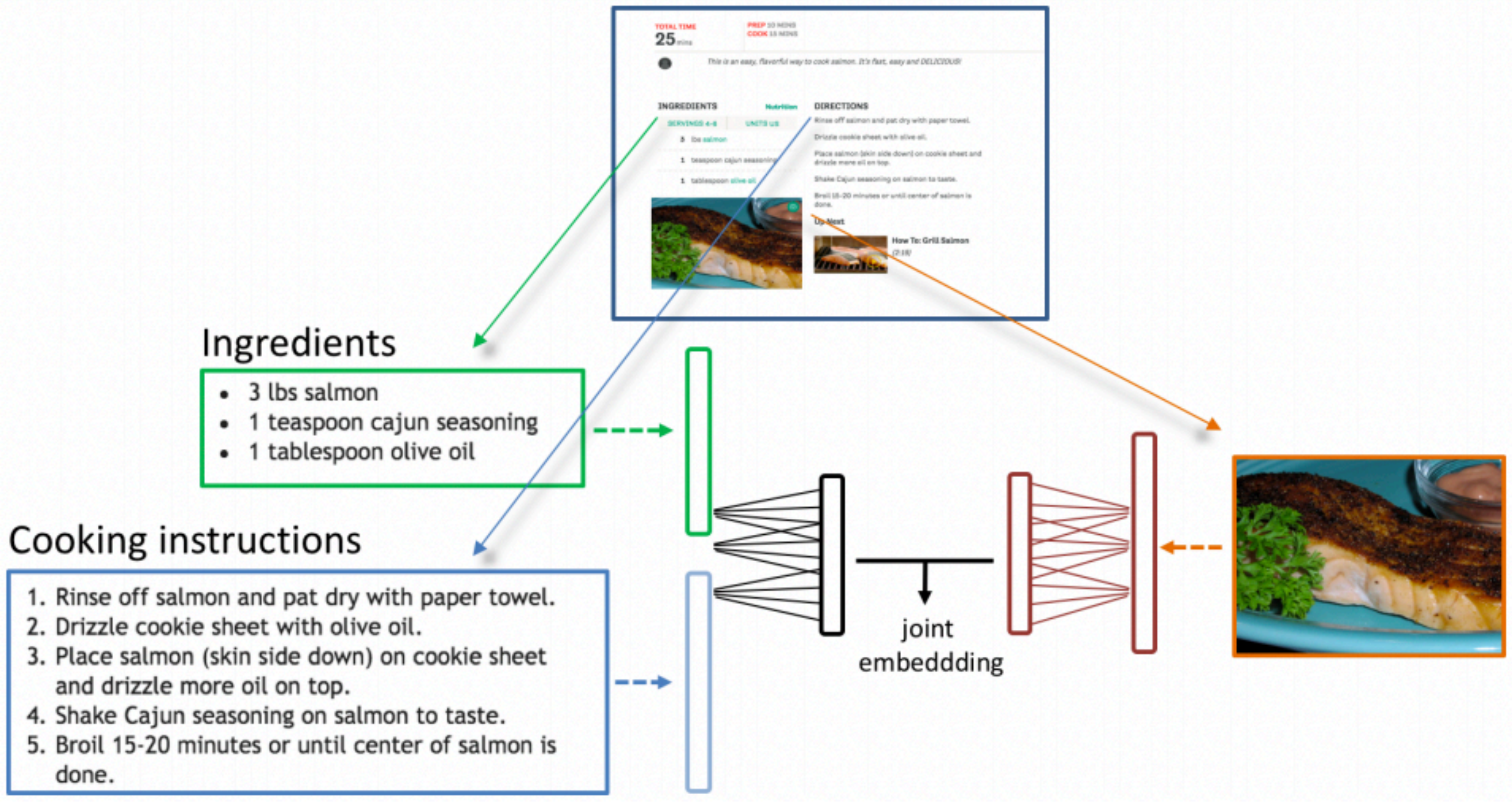
"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

"boy is doing backflip on wakeboard."

Context: Vision and Language



TOTAL TIME
25 mins

PREP 10 MINS
COOK 15 MINS

This is an easy, flavorful way to cook salmon. It's fast, easy and DELICIOUS!

INGREDIENTS Nutrition

SERVINGS 4-6 UNITS US

3 lbs salmon

1 teaspoon cajun seasoning

1 tablespoon olive oil

DIRECTIONS

Rinse off salmon and pat dry with paper towel.

Drizzle cookie sheet with olive oil.

Place salmon (skin side down) on cookie sheet and drizzle more oil on top.

Shake Cajun seasoning on salmon to taste.

Broil 15-20 minutes or until center of salmon is done.

Up Next

How To: Grill Salmon
(2:18)

Ingredients

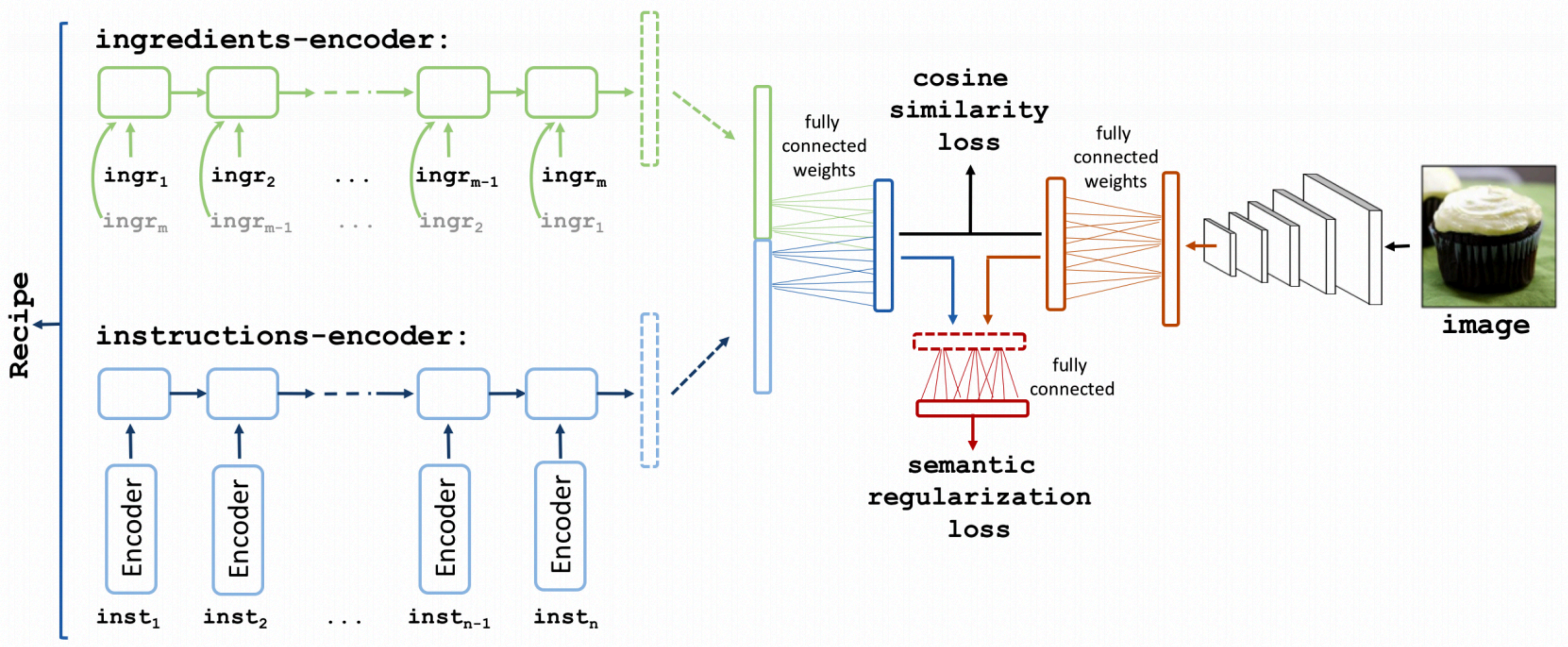
- 3 lbs salmon
- 1 teaspoon cajun seasoning
- 1 tablespoon olive oil

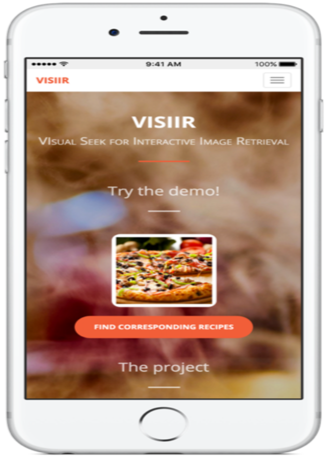
Cooking instructions

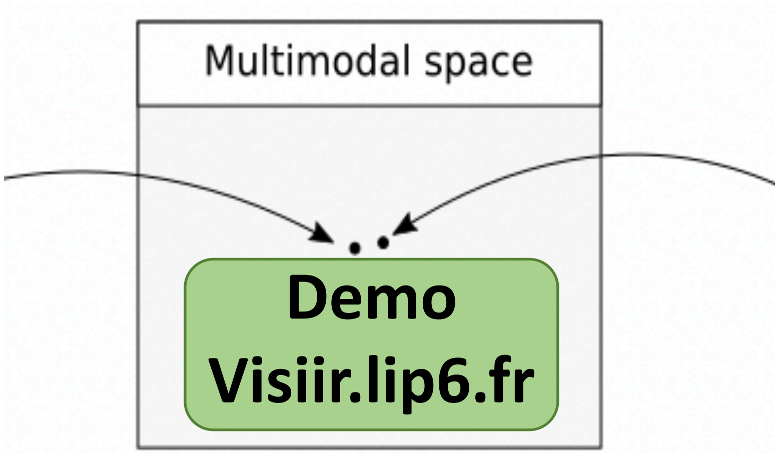
1. Rinse off salmon and pat dry with paper towel.
2. Drizzle cookie sheet with olive oil.
3. Place salmon (skin side down) on cookie sheet and drizzle more oil on top.
4. Shake Cajun seasoning on salmon to taste.
5. Broil 15-20 minutes or until center of salmon is done.

Learning Cross-modal Embeddings for Cooking Recipes and Food Images. A. Salvador,…, A. Torralba. Computer Vision and Pattern Recognition (CVPR), 2017

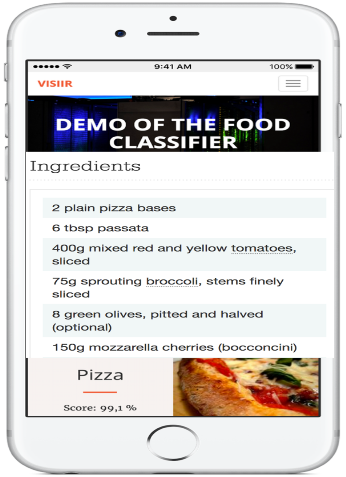
Context: Vision and Language




9:41 AM
100%
VISIIR
VISIIR
Visual Seek for Interactive Image Retrieval
Try the demo!
FIND CORRESPONDING RECIPES
The project



Demo
Visiir.lip6.fr



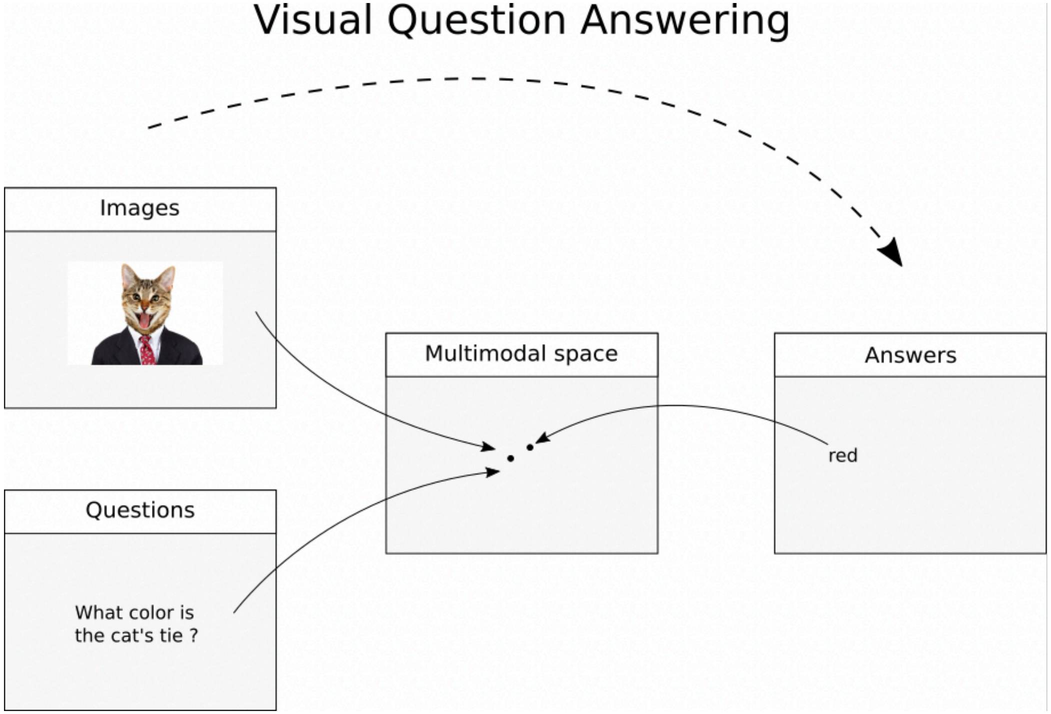
Ingredients

- 2 plain pizza bases
- 6 tbsp passata
- 400g mixed red and yellow tomatoes, sliced
- 75g sprouting broccoli, stems finely sliced
- 8 green olives, pitted and halved (optional)
- 150g mozzarella cherries (bocconcini)

Pizza

Score: 99,1 %

VQA



Visual Question Answering

Images

Multimodal space

Answers

red

Questions

What color is the cat's tie ?

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

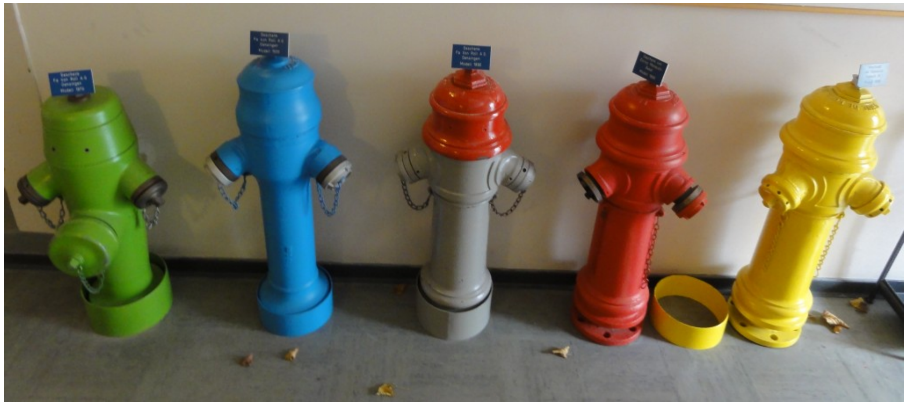
How many slices of pizza are there?
Is this a vegetarian pizza?

COCOQA 15756
What does the man rid 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)

VQA

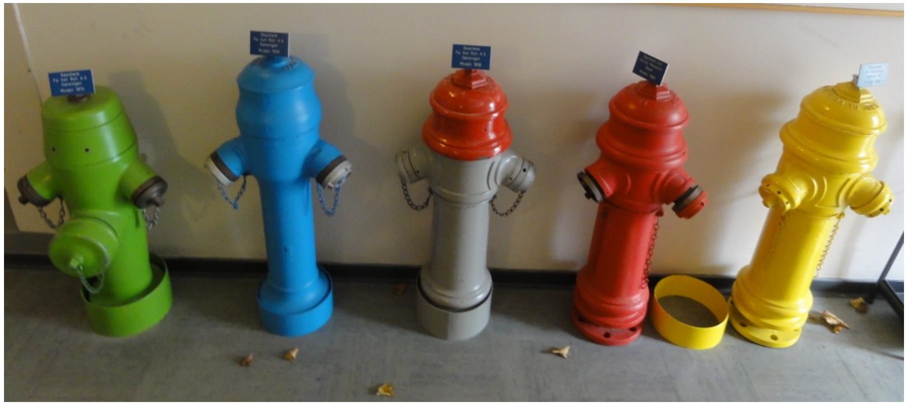
What color is the fire Hydrant on the left?



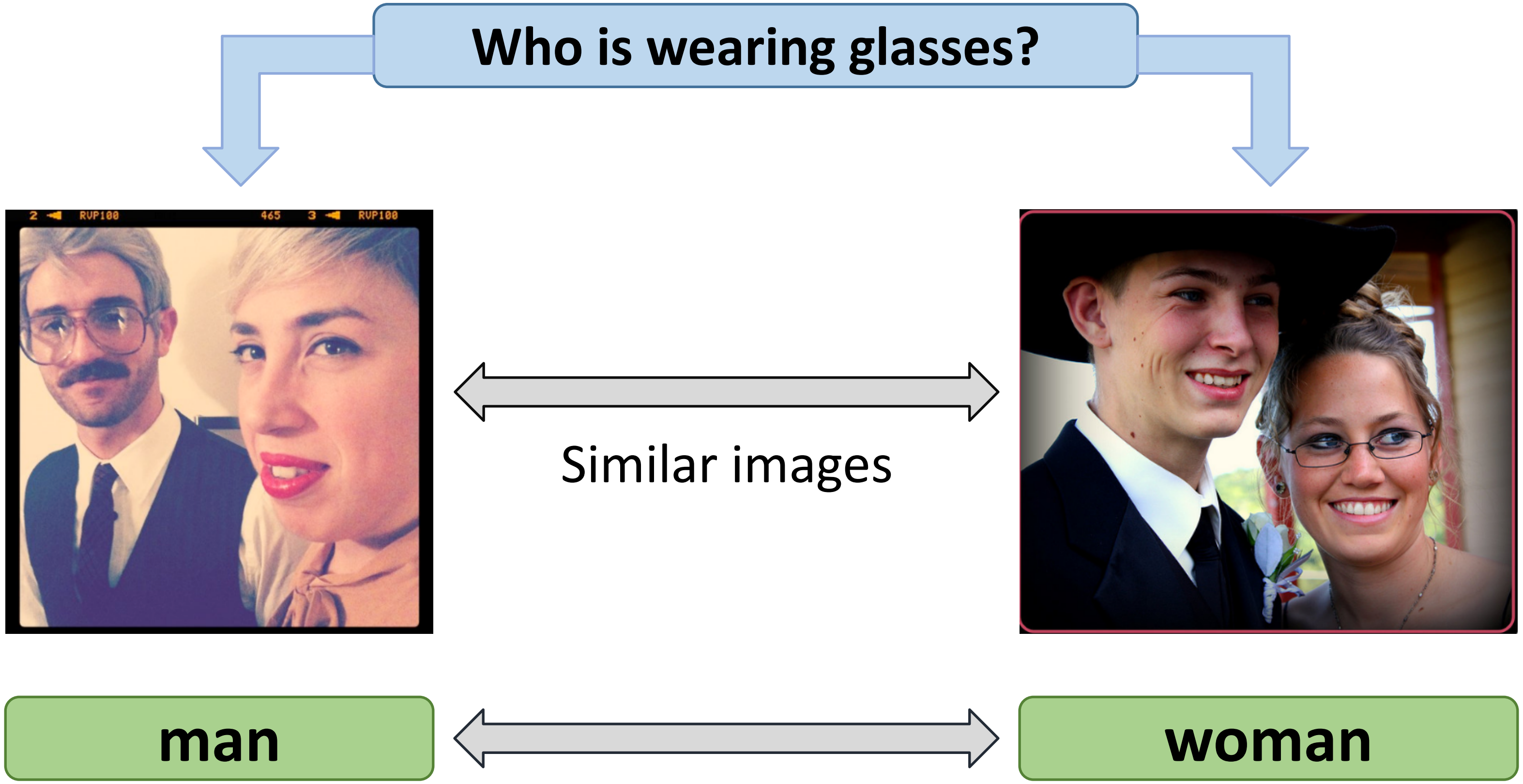
Green

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

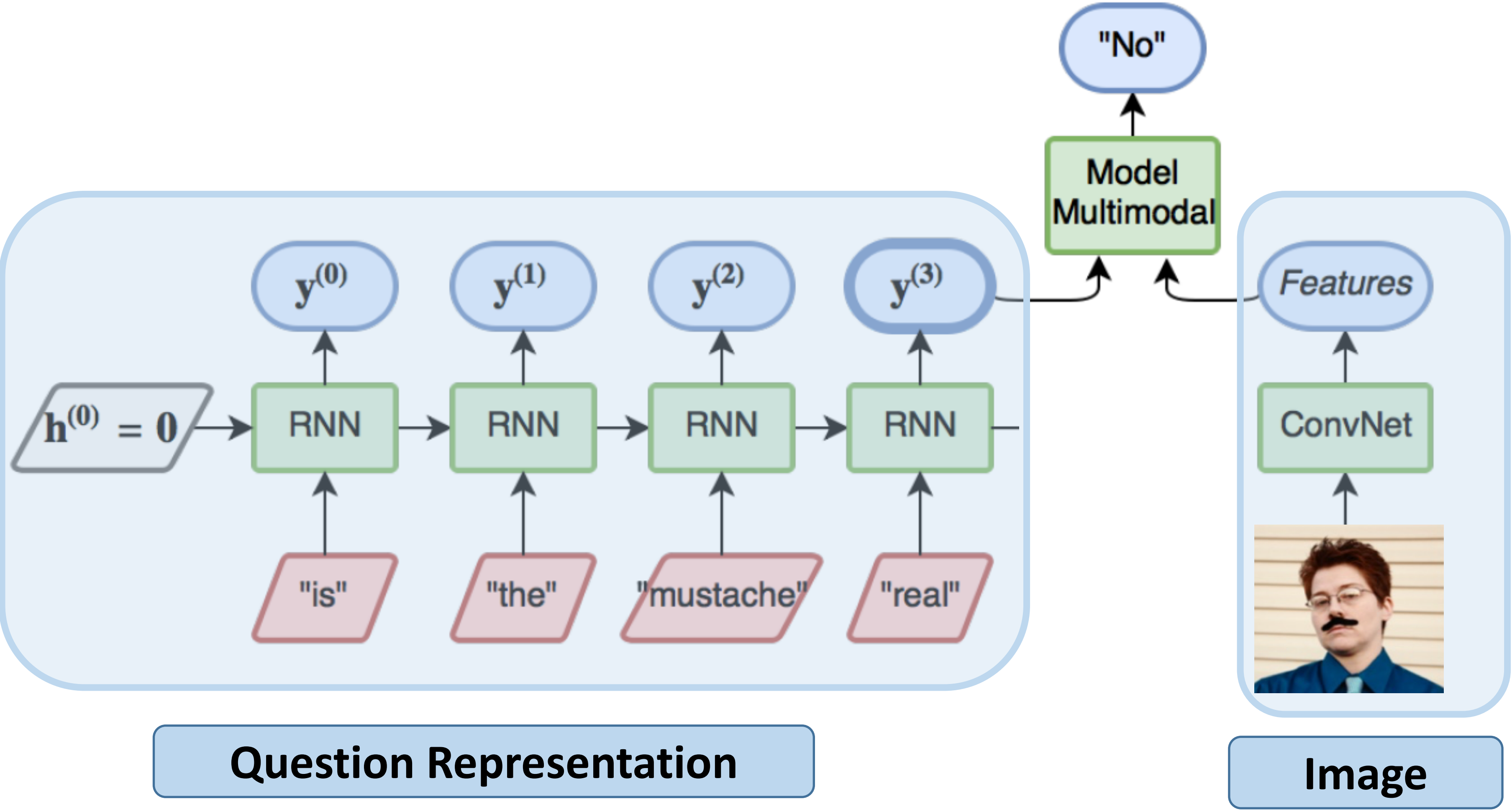
⇒ Need very good Visual and Question (deep) representations

- ⇒ Full scene understanding

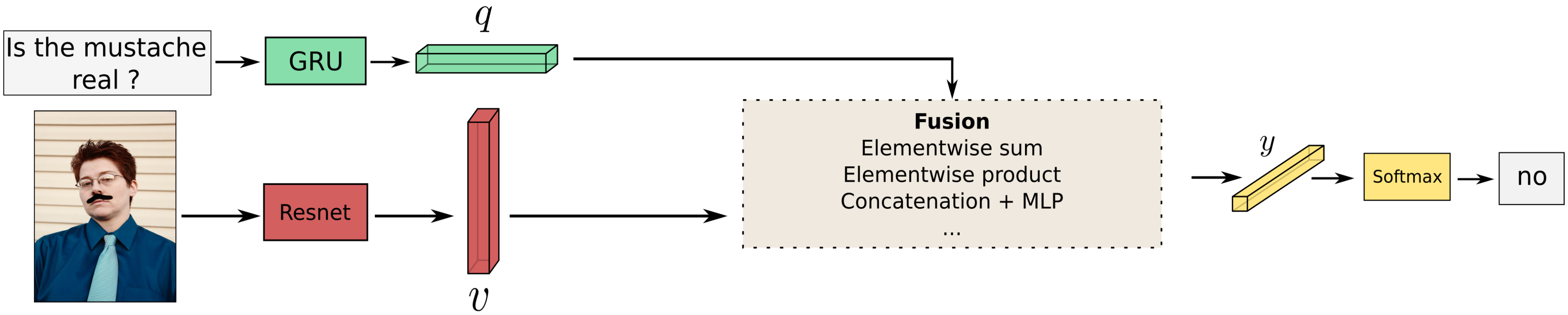
⇒ Need High level multimodal interaction modeling

- ⇒ Merging operators, attention and reasoning

Vanilla VQA scheme: 2 deep + fusion



VQA: fusion



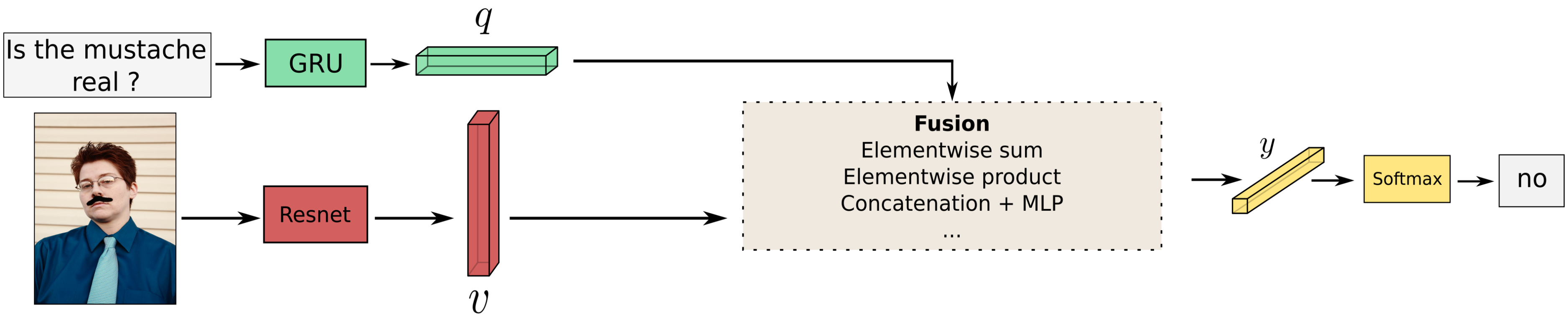
Concatenation & projection : $y = \boldsymbol{W} \begin{bmatrix} \mathbf{q} \\ \mathbf{v} \end{bmatrix}$

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

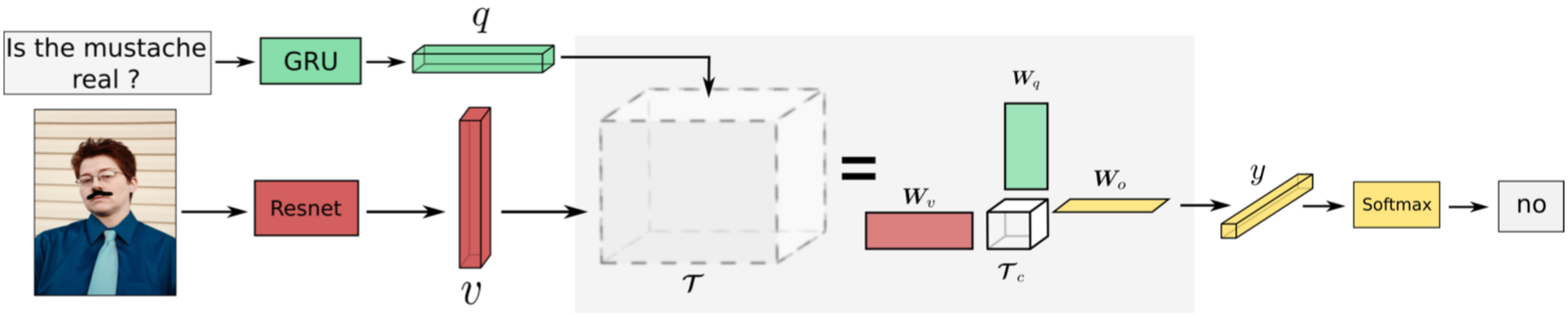
Element-wise sum : $y = (\boldsymbol{W}\mathbf{q}) + (\boldsymbol{V}\mathbf{v})$

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: 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**

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

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

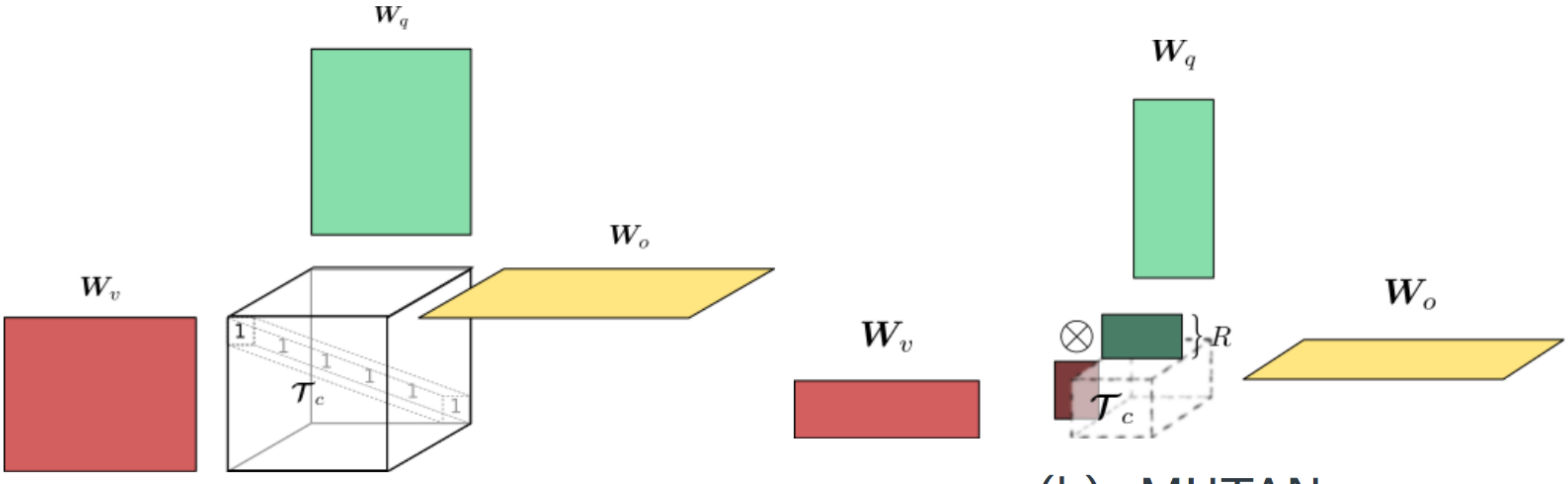
VQA: fusion

Learn the 3-ways Tensor coeff.

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

Need to reduce the Tensor Size:

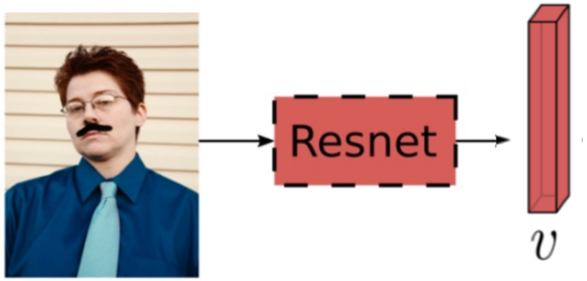
- Tucker based decomposition



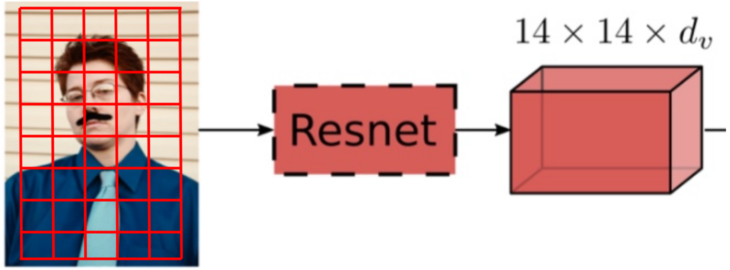
(a) MLB

(b) MUTAN

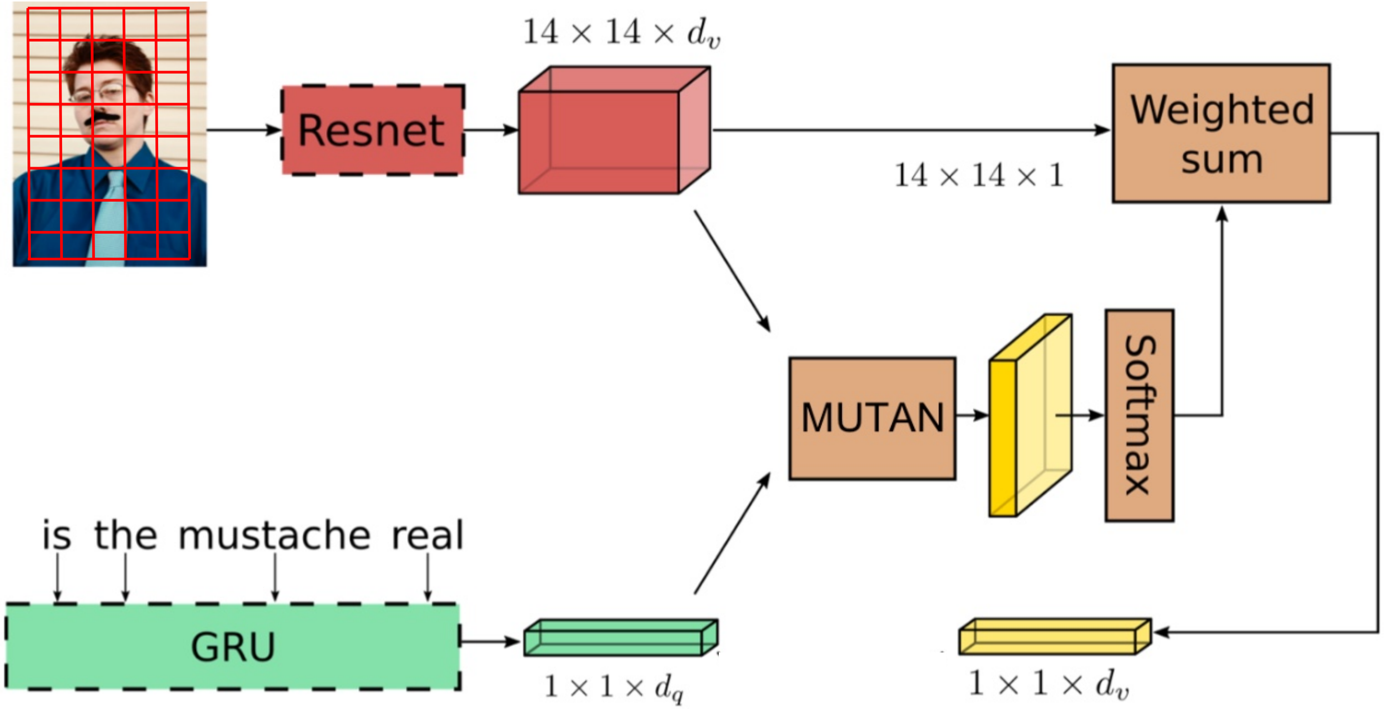
VQA: Attention process & reasoning



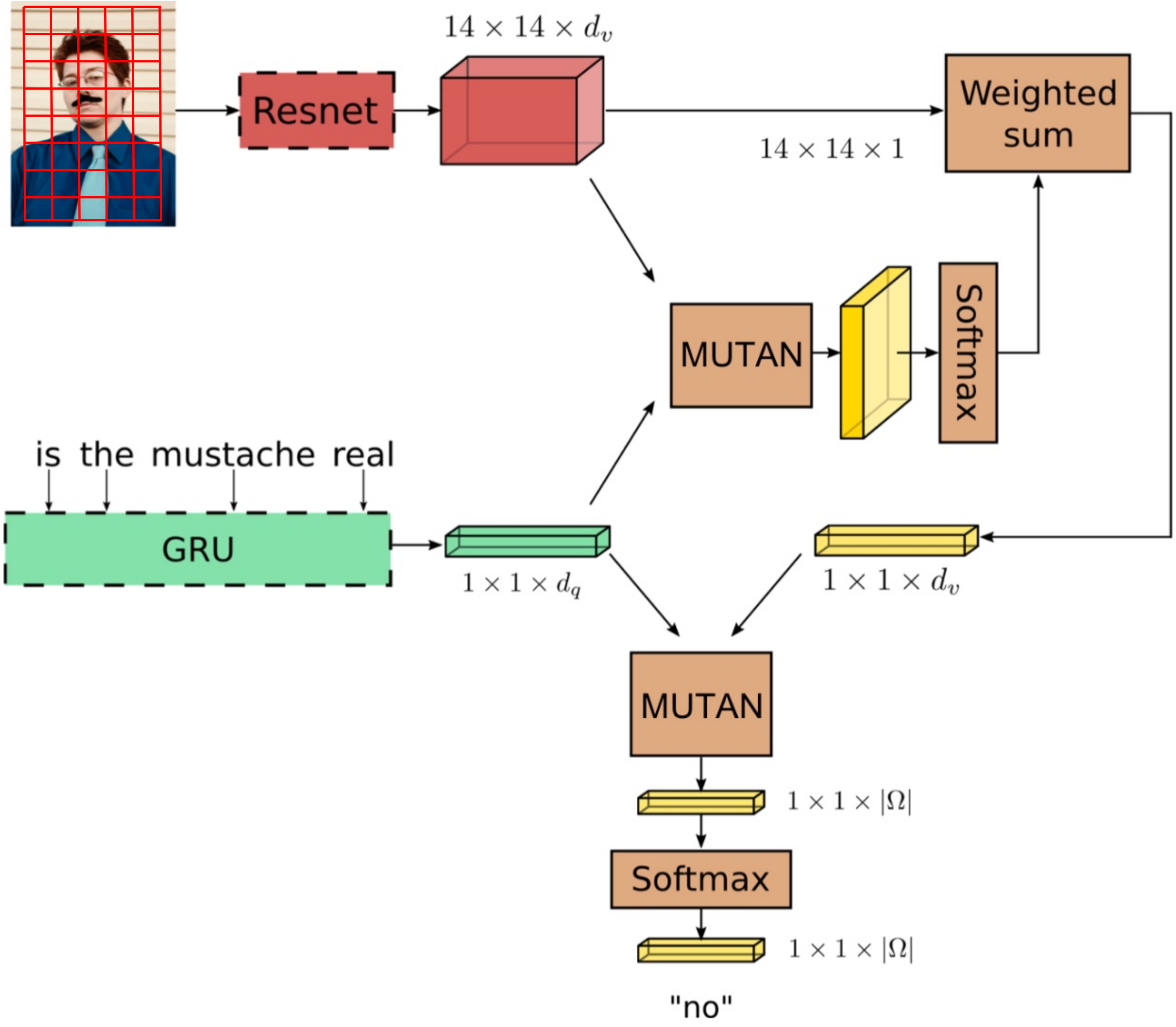
VQA: Attention process & reasoning



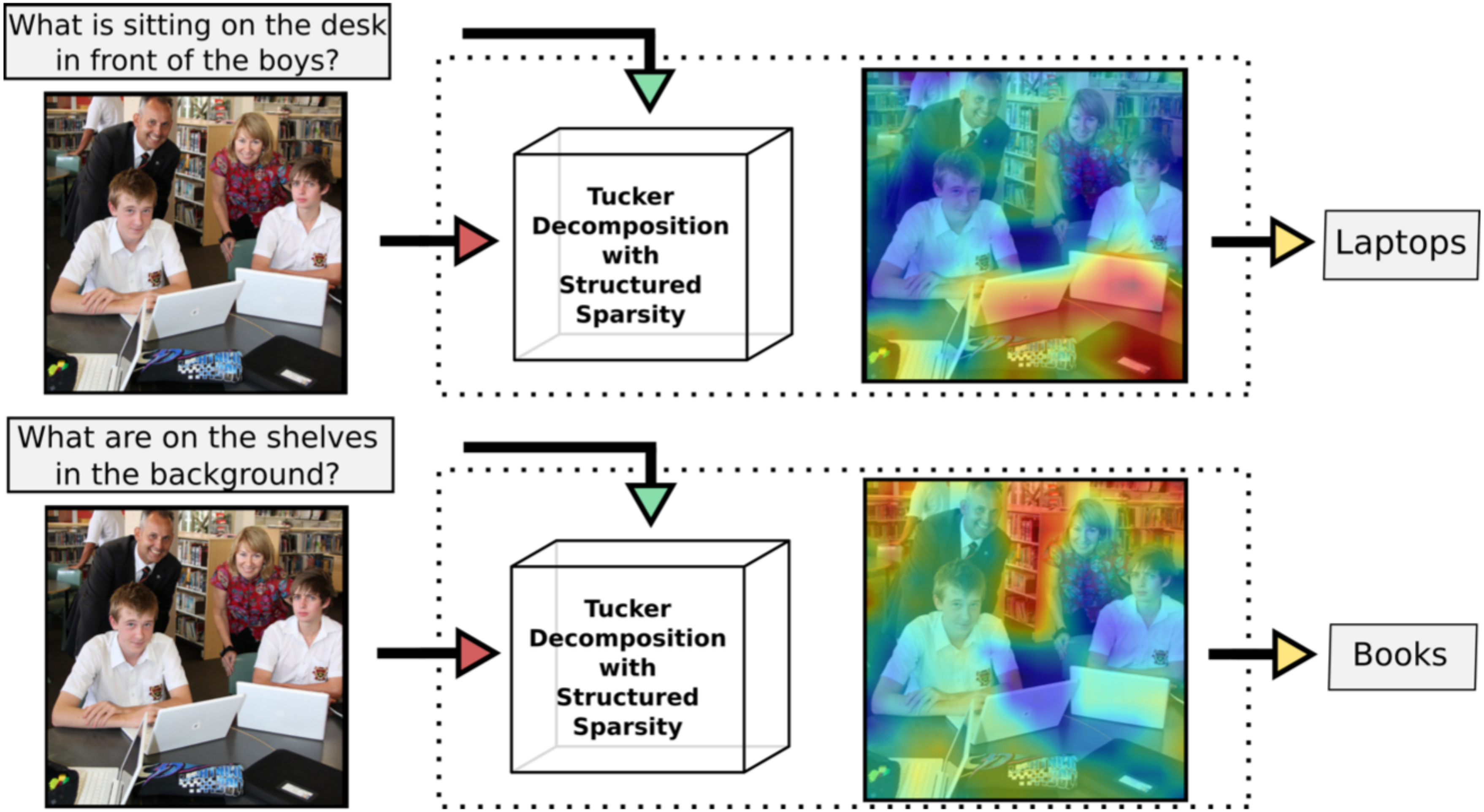
VQA: Attention process & reasoning



VQA: Attention process & reasoning



VQA: Attention process & reasoning



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

Tucker Decomposition with Structured Sparsity

Laptops

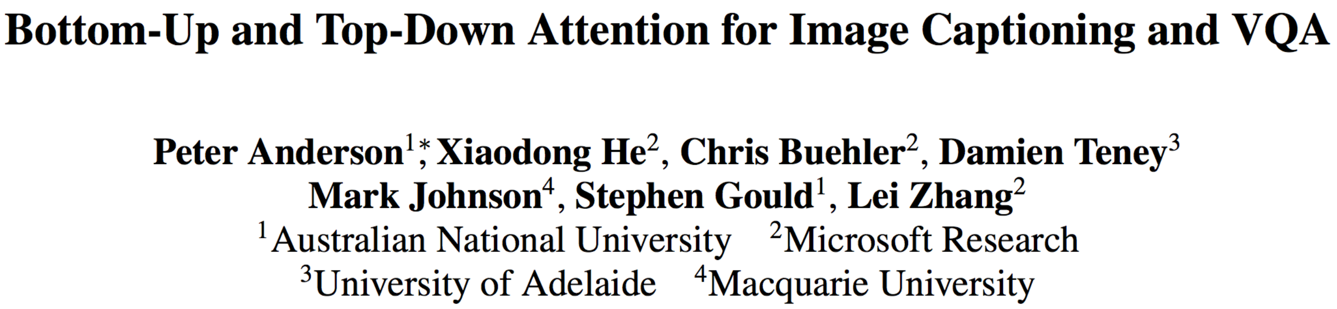
What are on the shelves in the background?

Tucker Decomposition with Structured Sparsity

Books

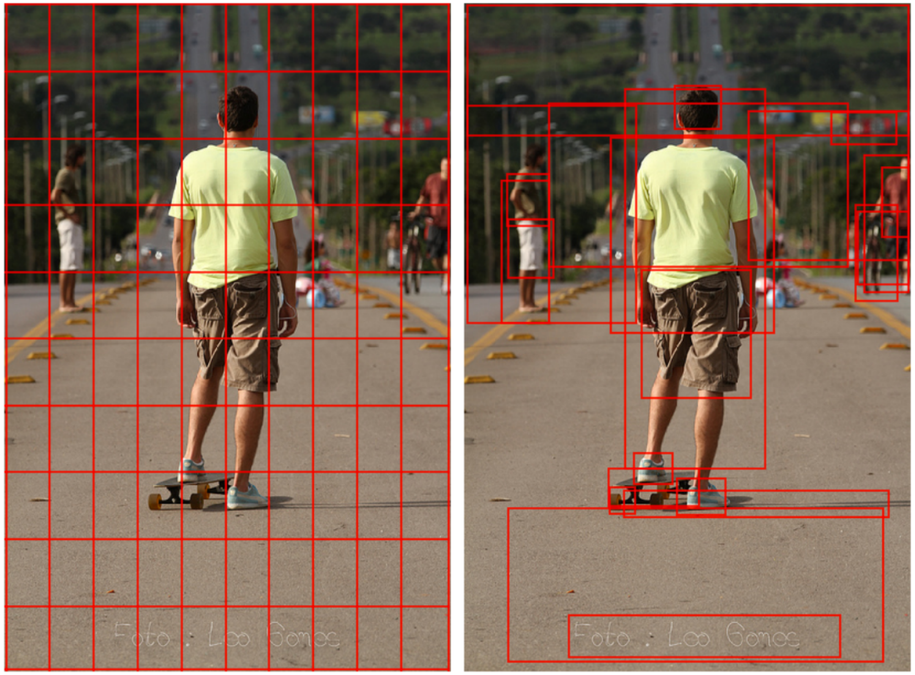
VQA: Attention process & reasoning

Winner of the VQA Challenge in CVPR 2017:



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]
[1]Australian National University [2]Microsoft Research
[3]University of Adelaide [4]Macquarie University



VQA: Attention process & reasoning

Many initiatives to improve datasets and evaluate reasoning as:

VQA v2.0 dataset and challenge 2017

- [Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering, Y. Goyal, **D. Batra, D. Parikh**, CVPR 2017]

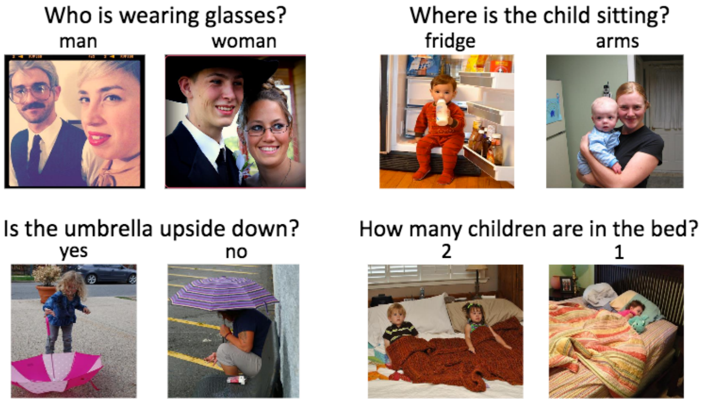
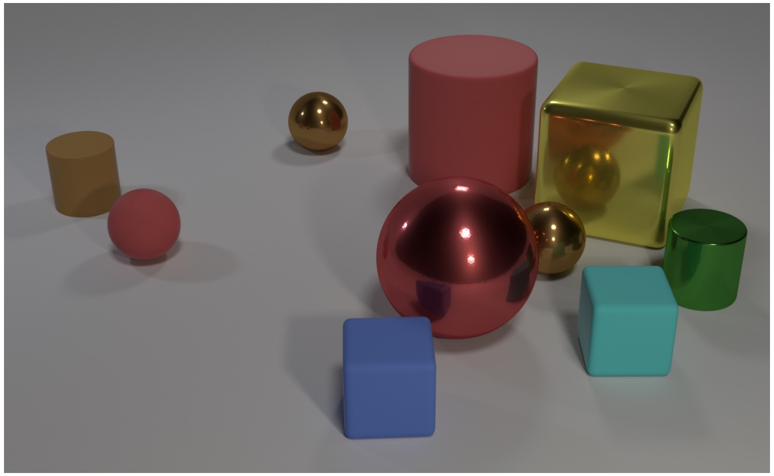

Who is wearing glasses?
man
woman
Where is the child sitting?
fridge
arms
Is the umbrella upside down?
yes
no
How many children are in the bed?
2
1

Figure 1: Examples from our balanced VQA dataset.

[CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning, Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, Ross Girshick, CVPR 2017]

- Questions testing various aspects of visual reasoning including **attribute identification**, **counting**, **comparison**, **spatial relationships**, and **logical operations**.

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



MLIA/Chordettes team: Matthieu Cord http://webia.lip6.fr/~cord
D.Picard (CNRS delegation), PhD T. Durand, T. Robert, T. Mordan, X. Wang, M. Blot, M. Carvahlo, H. BenYounes, R. Cadene, Y. Chen, E. Mehr, M. Engilberge, D. Brooks;
Collab. N. Thome (CNAM), P. Perez (TECHNICOLOR)

MUTAN: Multimodal Tucker Fusion for Visual Question Answering
H. Ben-Younes*, R. Cadene*, N. Thome, M. Cord, ICCV (2017) (*equal contrib.)
Pytorch code: https://github.com/Cadene

Our Deep Recipe Reco on your mobile: visiir.lip6.fr

Recent refs. on Deep learning for Visual Recognition

- Deformable Part-based Fully Convolutional Network for Object Detection, T. Mordan, N. Thome, M. Cord, G. Henaff, BMVC 2017 (**Best paper**)
- WILDCAT: Weakly Supervised Learning of Deep ConvNets for Image Classification, Pointwise Localization and Segmentation, T. Durand, T. Mordan, N. Thome, M. Cord, CVPR 2017
- WELDON: Weakly Supervised Learning of Deep Convolutional Neural Networks, T. Durand, N. Thome, M. Cord, CVPR 2016
- Deep Neural Networks Under Stress, M. Carvalho, M. Cord, S. Avila, N. Thome, E. Valle, ICIP 2016
- LR-CNN for fine-grained classification with varying resolution, M Chevalier+, ICIP 2015
- Learning Deep Hierarchical Visual Feature Coding, H. Goh+, IEEE TNNLS 2014
- Sequentially generated instance-dependent image representations for classification, G Dulac-Arnold, L Denoyer, N Thome, M Cord, P Gallinari, ICLR 2014
- Top-Down Regularization of Deep Belief Networks, H. Goh, N. Thome, M. Cord, JH. Lim, NIPS 2013