



Visual Question Answering: a new task for Vision and Language understanding

Matthieu Cord
Sorbonne Université
Computer science dpt. LIP6
MLIA
UMR CNRS

Outline

1. Context: Vision and Language
2. Multimodal embedding
3. VQA framework

Visual Question Answering

Question Answering:

Is Patrick in the room?

Visual Question Answering

Visual Question Answering:

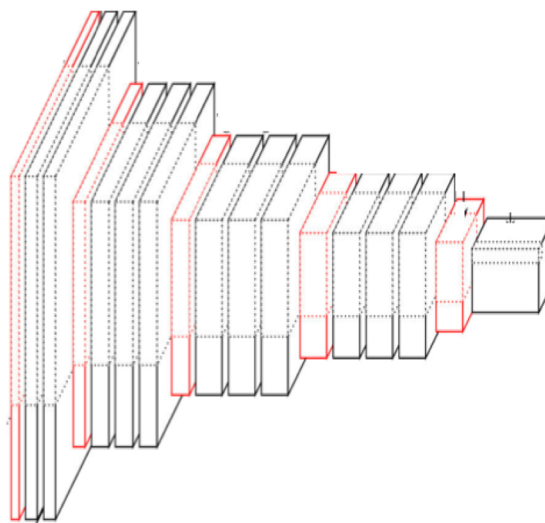
Is Patrick in the room?



Visual Question Answering

Visual Question Answering:

Is Patrick in the room?



Yes/No

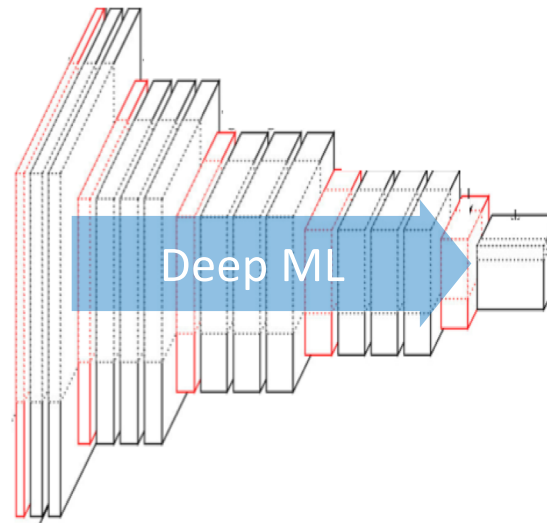
On the left

At the back ...

Visual Question Answering

Visual Question Answering:

Is Patrick in the room?



Yes/No

On the left

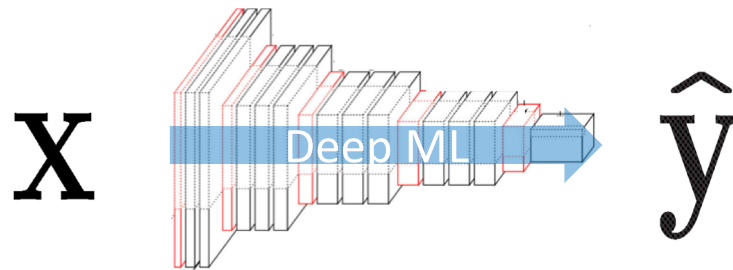
At the back ...

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

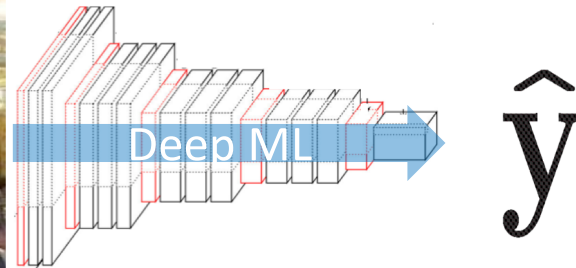
Classification: from Image to keywords/labels



Available Web demo (@Clarifai)

Context: Vision and Language

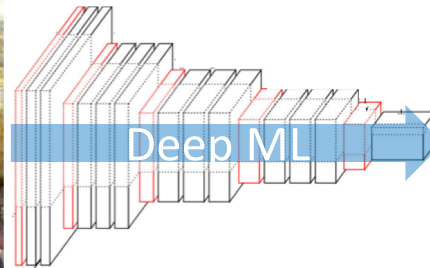
Classification: from Image to keywords/labels



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.

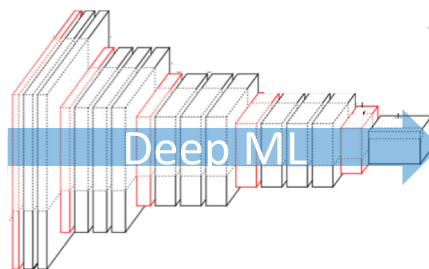
Le Monde.fr | 13.11.2015 à 11h40 • Mis à jour le 13.11.2015 à 16h25 |

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)

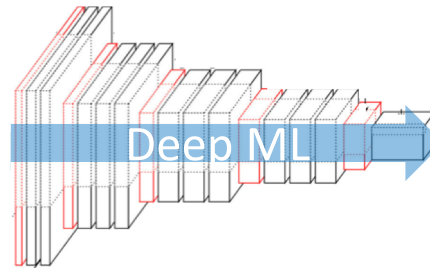
Context: Vision and Language

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)

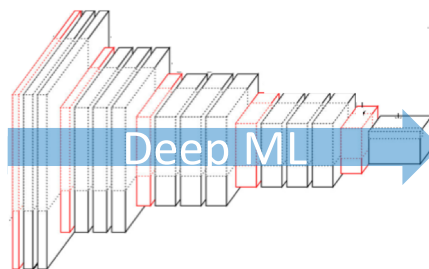
Context: Vision and Language

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



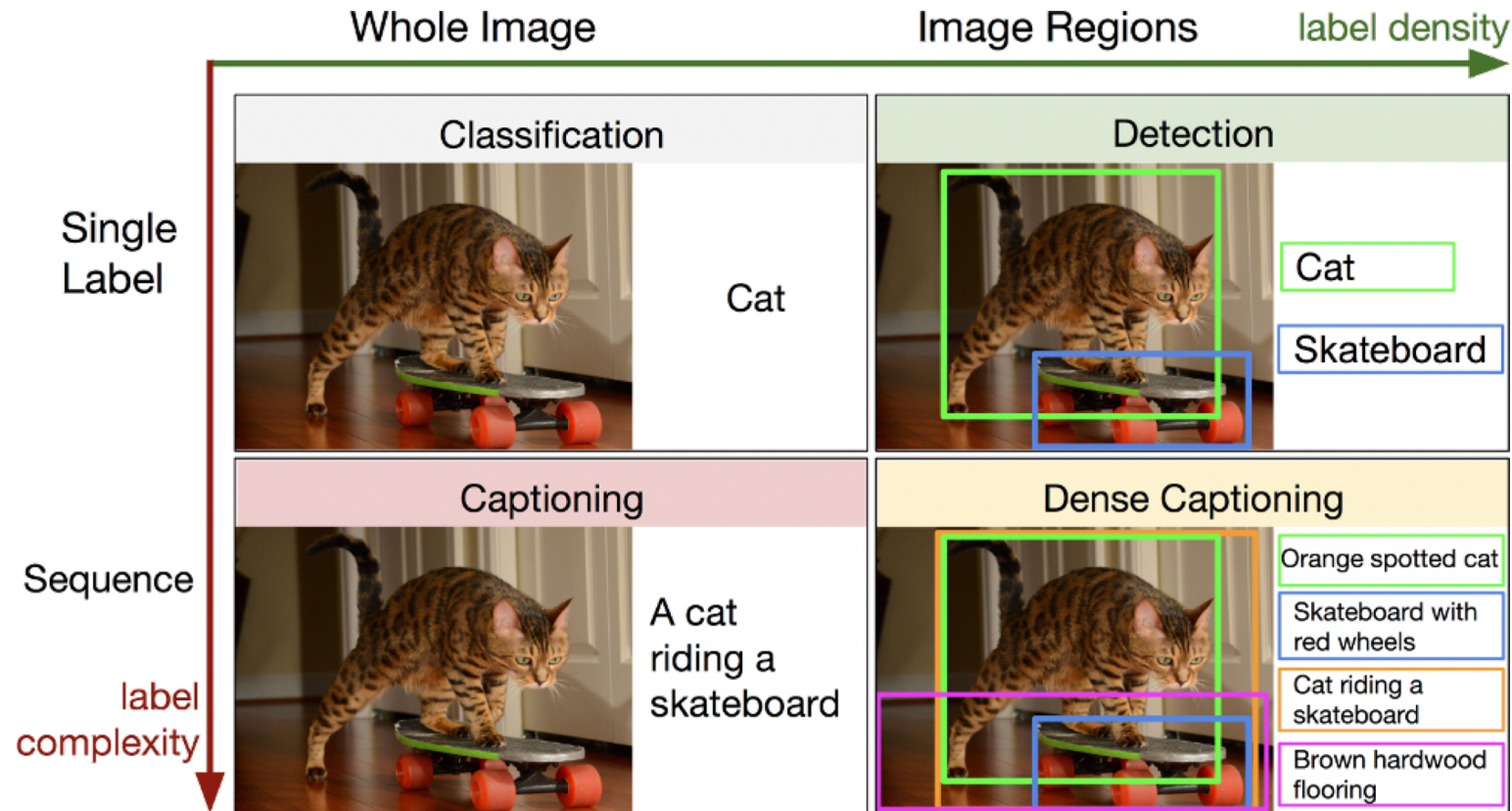
WILDCAT: Weakly Supervised Learning of Deep ConvNets..., T. Durand, T. Mordan, N. Thome, M. Cord, CVPR 2017

Deep ML for object localization: from pixel to labels



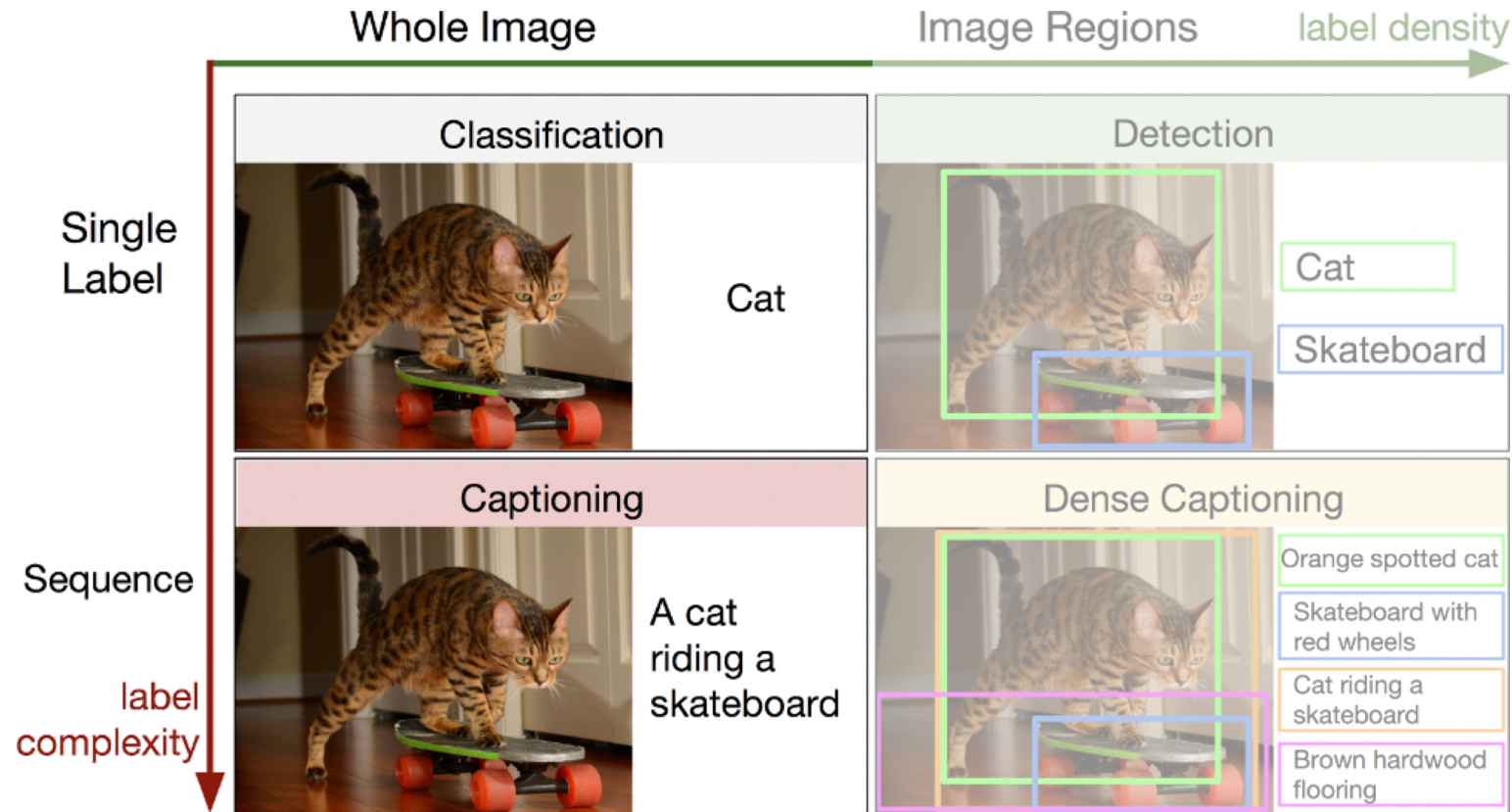
WILDCAT: Weakly Supervised Learning of Deep ConvNets..., T. Durand, T. Mordan, N. Thome, M. Cord, CVPR 2017

Context: Vision and Language



@Feifei

Context: Vision and Language



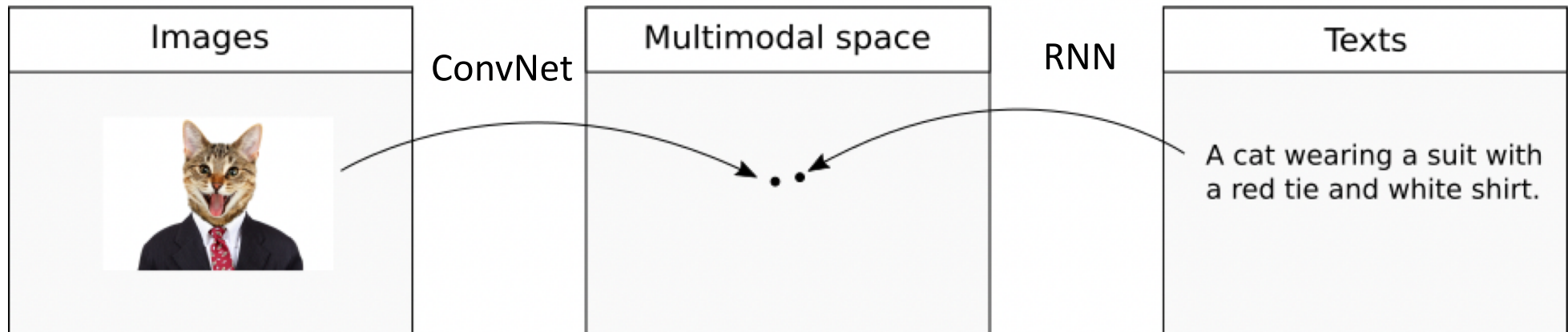
Language
description/complexity

Vision and Language: from keywords to sentence ...

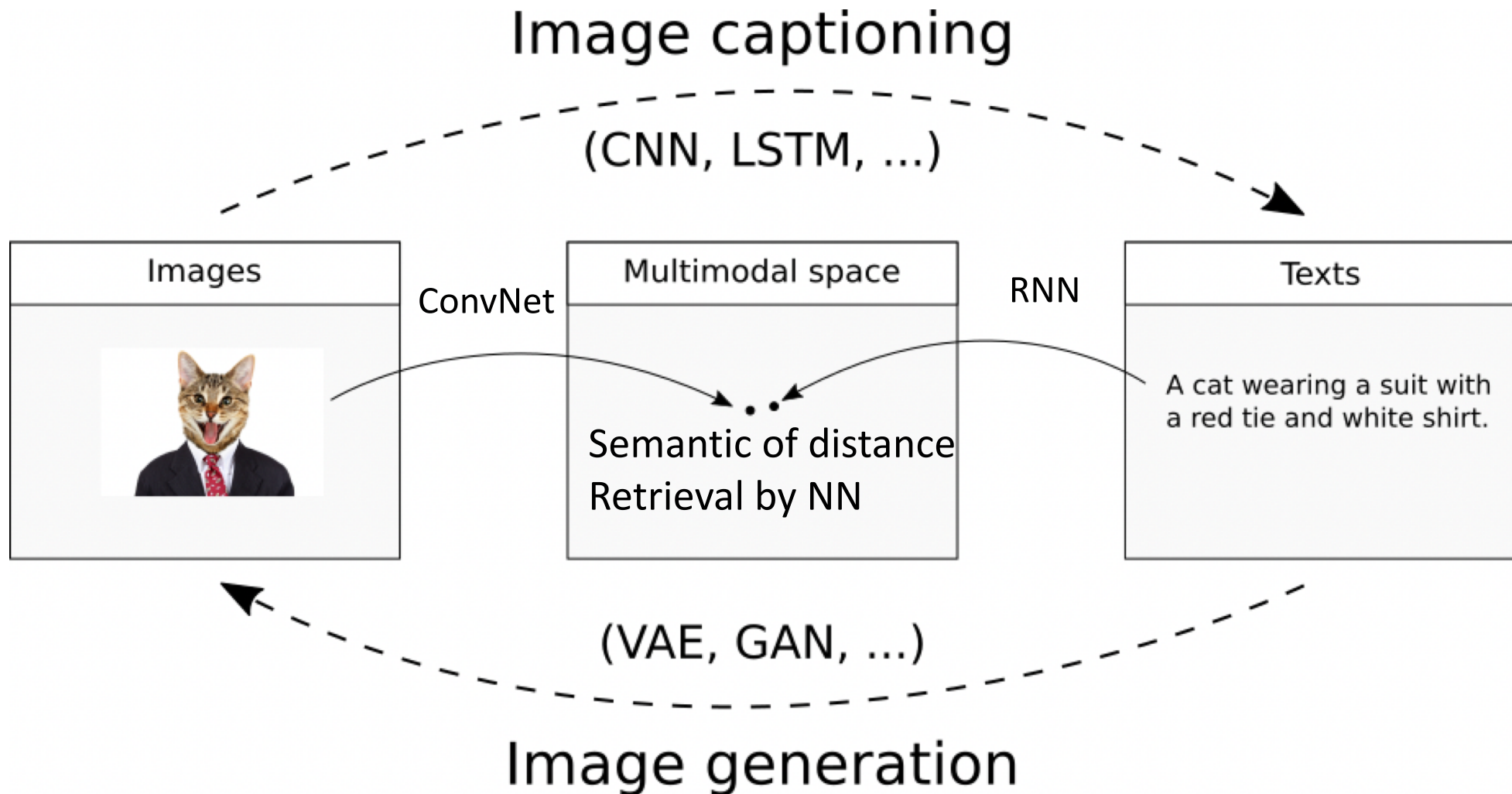
Outline

1. Context: Vision and Language
- 2. Multimodal embedding**
 - 1. Deep nets to align text+image**
 - 2. learning**
3. VQA framework
 1. Fusion in VQA
 2. Reasoning in VQA

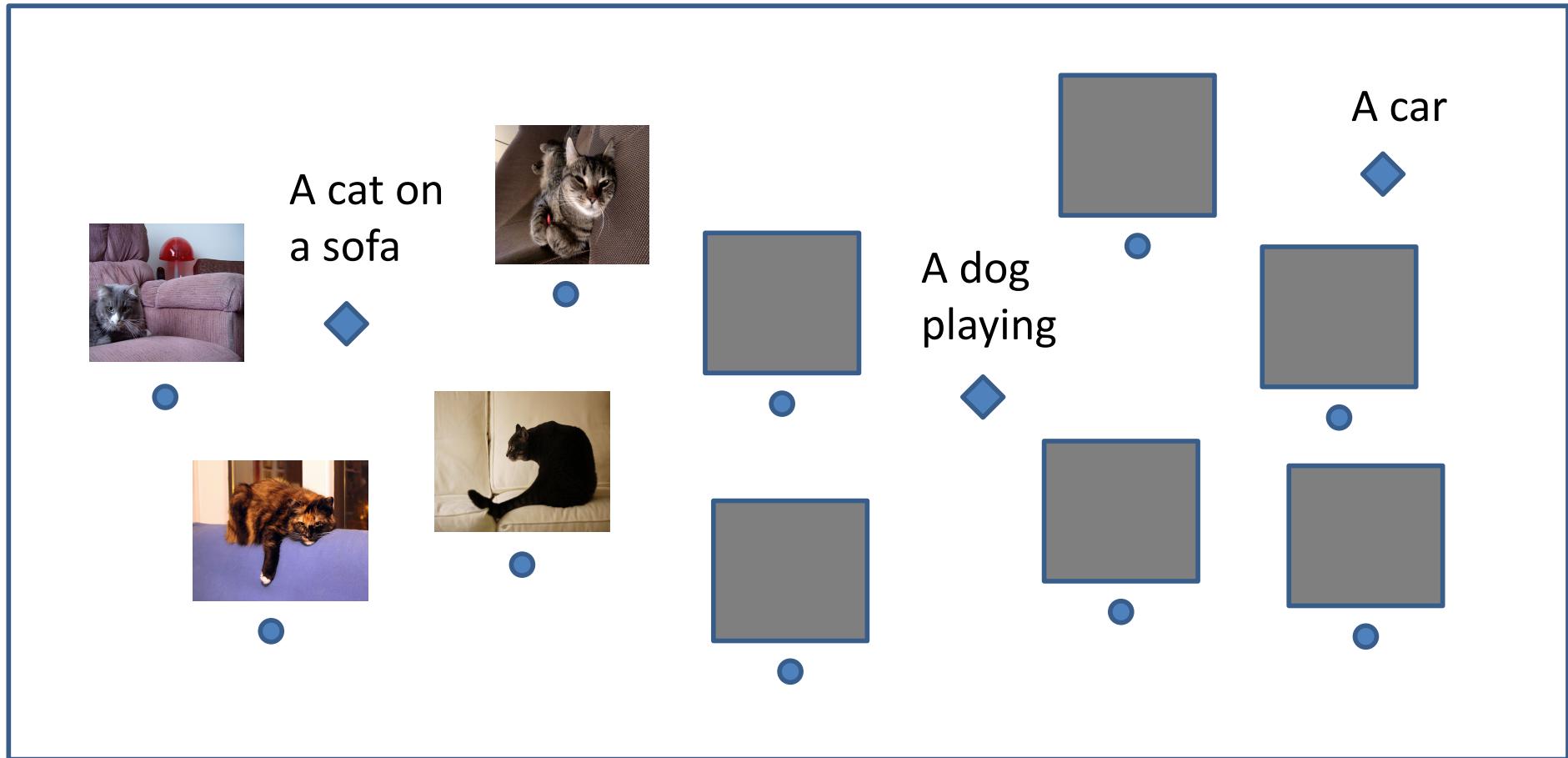
Deep semantic-visual embedding



Deep semantic-visual embedding



Deep semantic-visual embedding

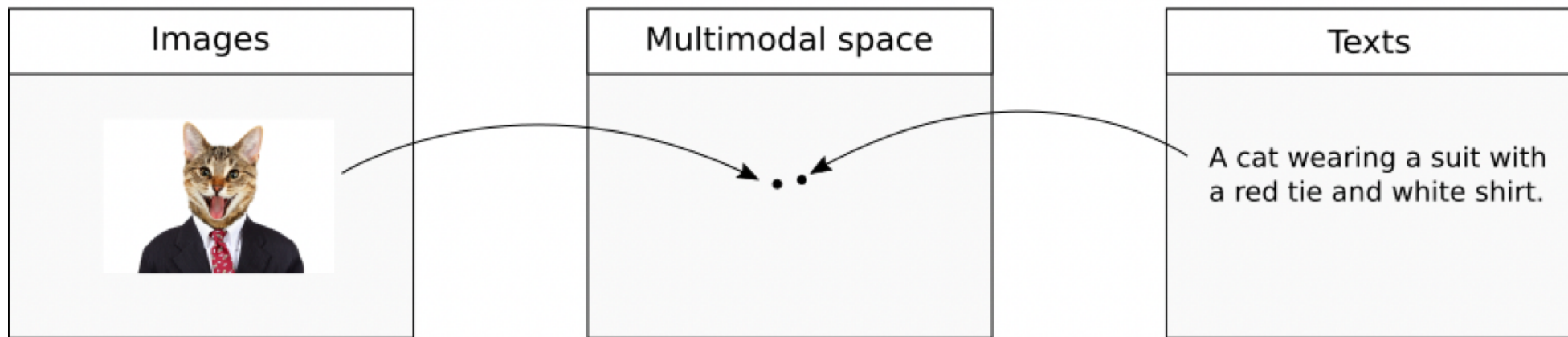


2D Semantic visual space example:

- 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

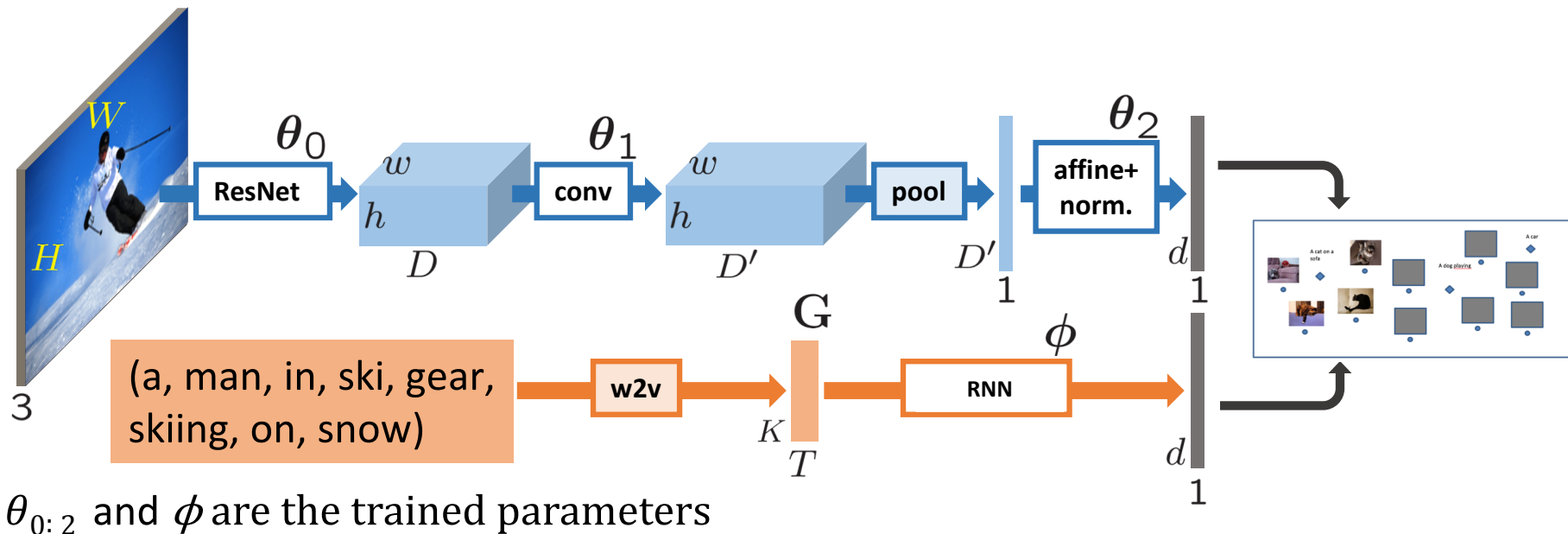
Finding beans in burgers: Deep semantic-visual embedding with localization,
M. Engilberge et al, CVPR 2018

Visual pipeline:

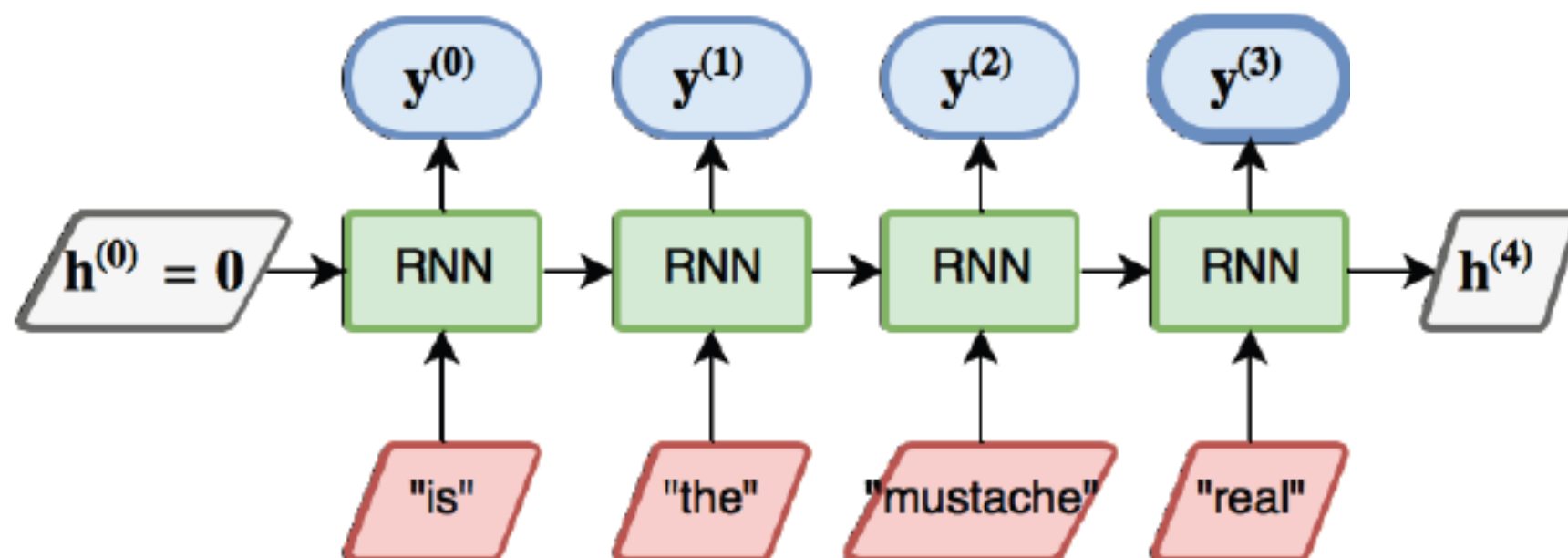
- ResNet-152 pretrained
- Weldon spatial pooling

Textual pipeline:

- Pretrained word embedding
- Simple Recurrent Unit (SRU)



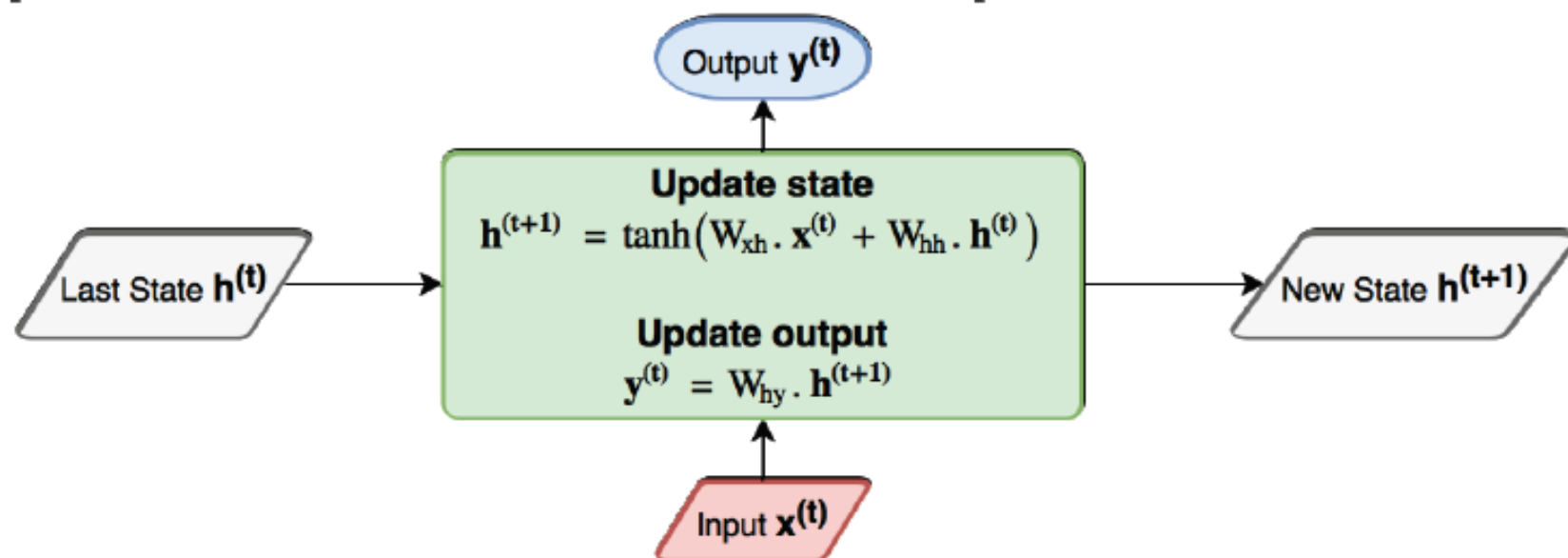
RNN



Output $\mathbf{y}^{(i)}$ is a vector of 2400 dimensions.
Consider the last output $\mathbf{y}^{(4)}$ to be the final output.

Vanilla RNN

[Goodfellow, Bengio, and Courville 2016]

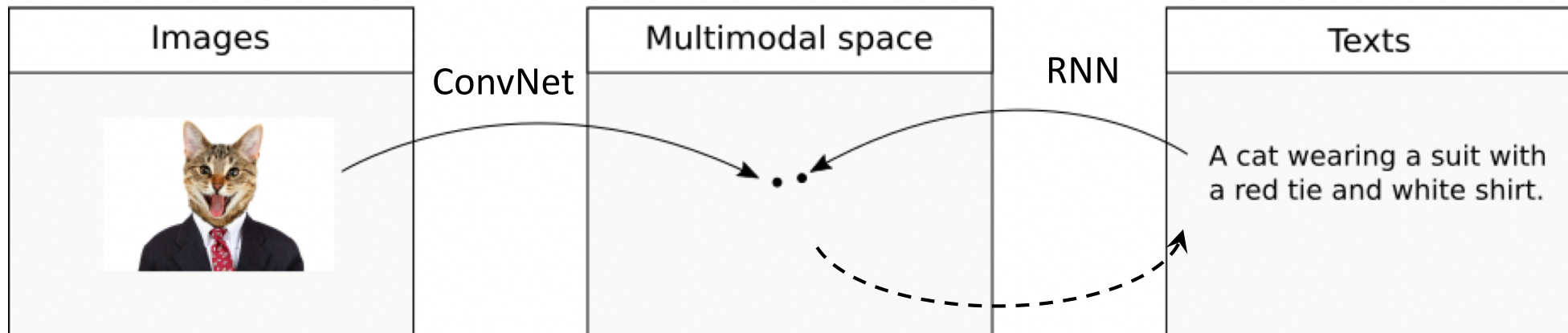


- Parameters matrices (weight) : \mathbf{W}_{xh} , \mathbf{W}_{hh} and \mathbf{W}_{hy}
- Parameters vectors (bias): \mathbf{b}_{xh} , \mathbf{b}_{hh} and \mathbf{b}_{hy}

Some results

Image captioning

(CNN, LSTM, ...)



"man in black shirt is playing guitar."



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



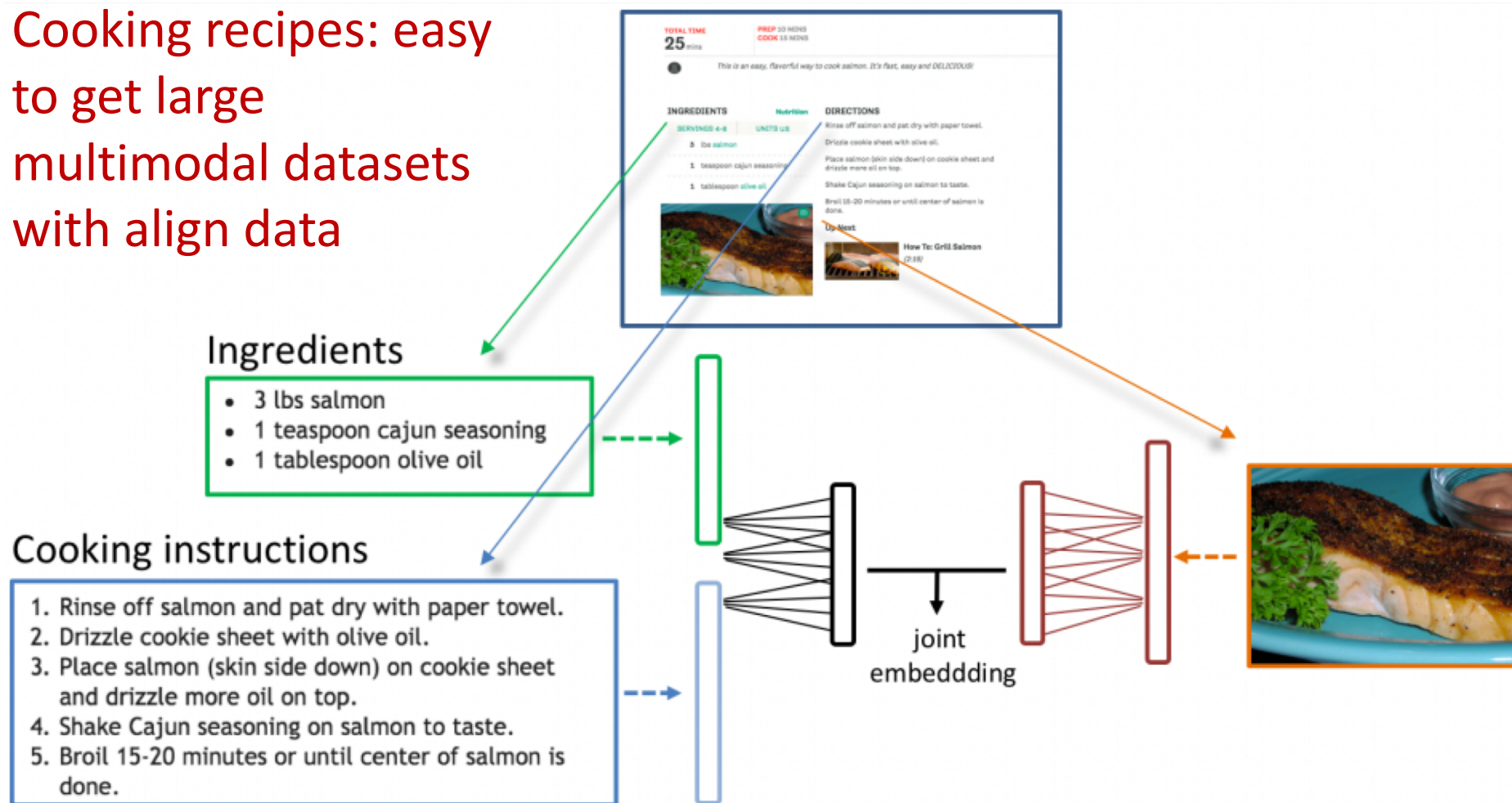
"two young girls are playing with lego toy."



"boy is doing backflip on wakeboard."

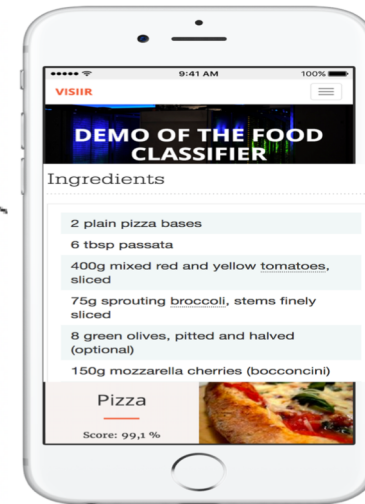
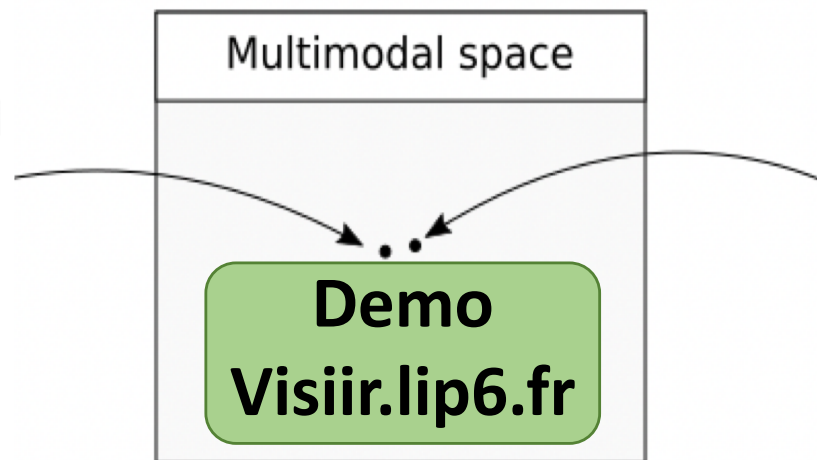
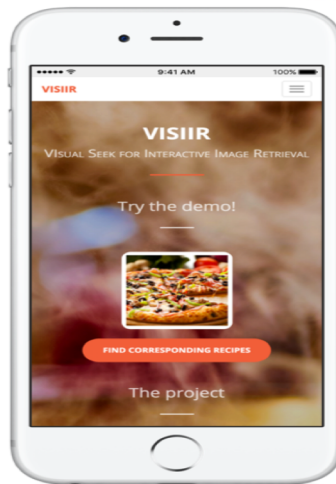
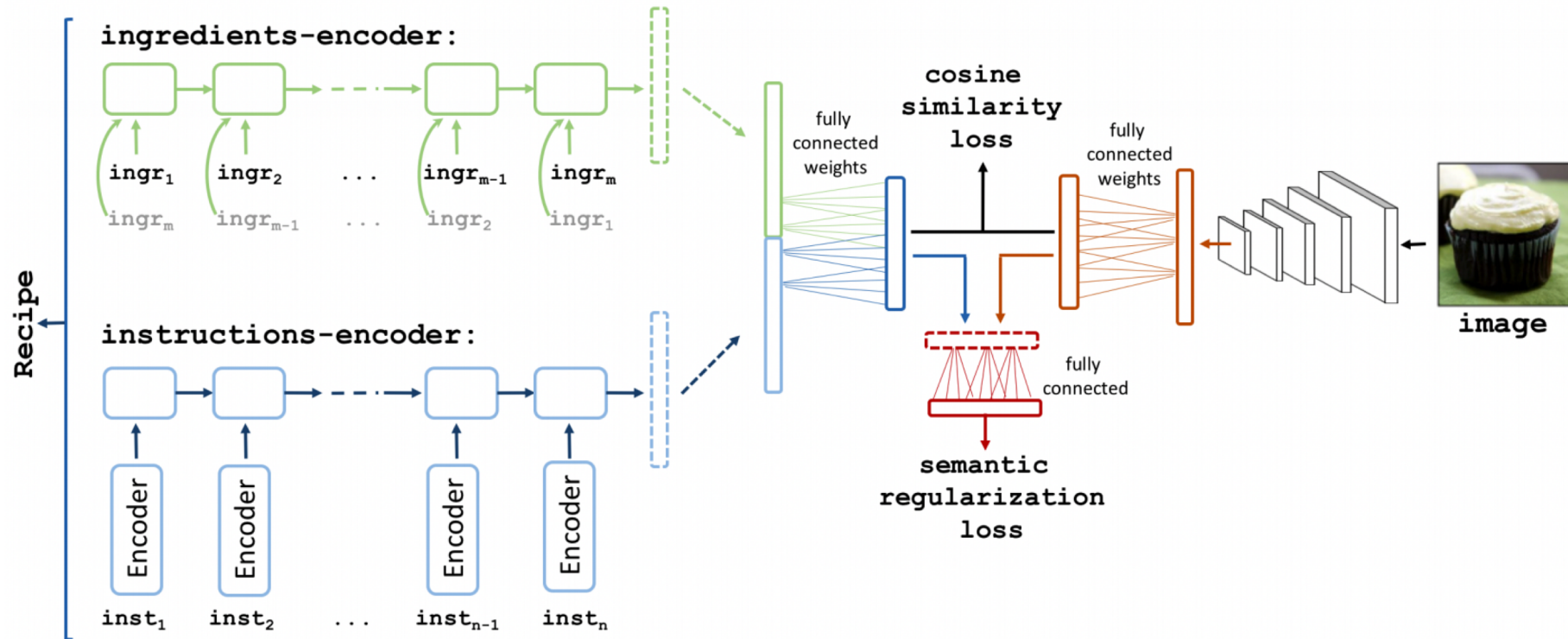
How to get large training datasets?

Cooking recipes: easy to get large multimodal datasets with align 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

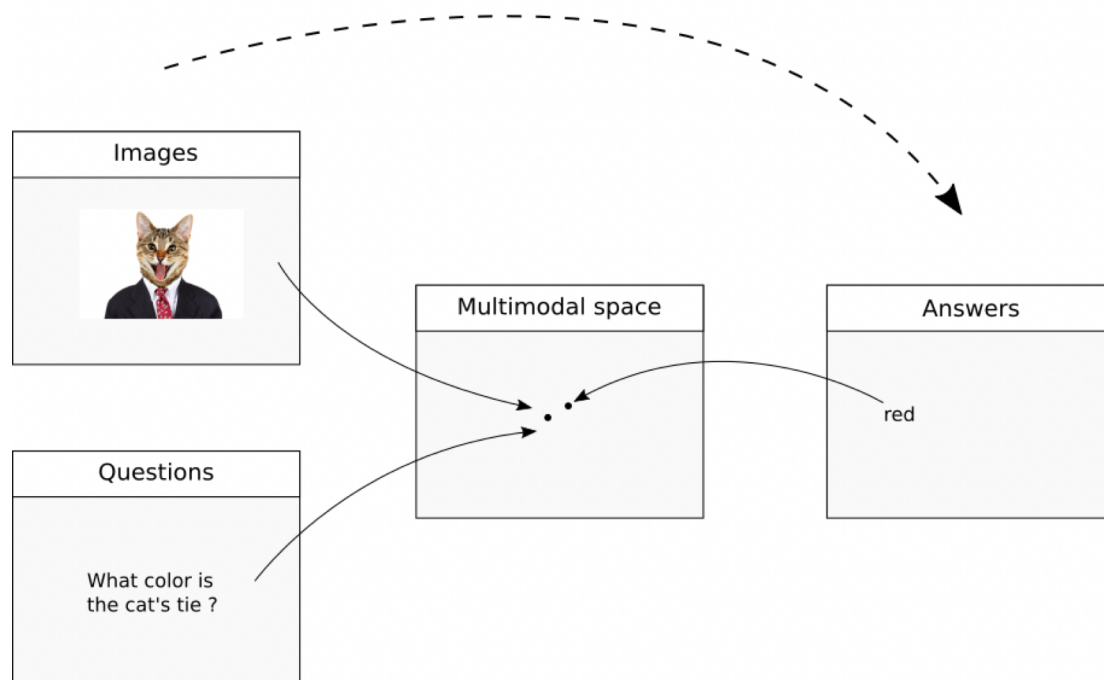


Outline

1. Context: Vision and Language
2. Multimodal embedding
 1. Deep nets to align text+image
 2. Learning
- 3. Visual Question Answering**
 1. Task modeling
 2. Fusion in VQA
 3. Reasoning in VQA

VQA

Visual Question Answering



COCOQA 15756

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)



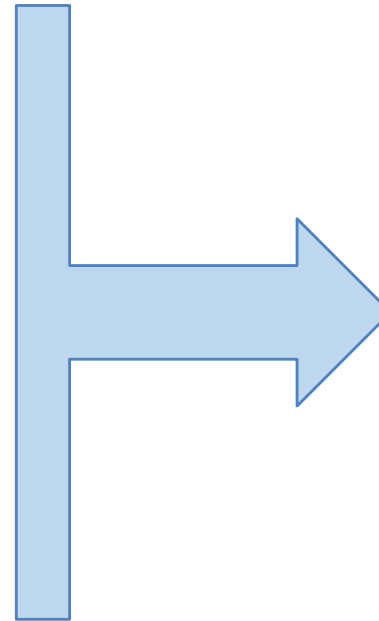
Does it appear to be rainy?
Does this person have 20/20 vision?



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

VQA

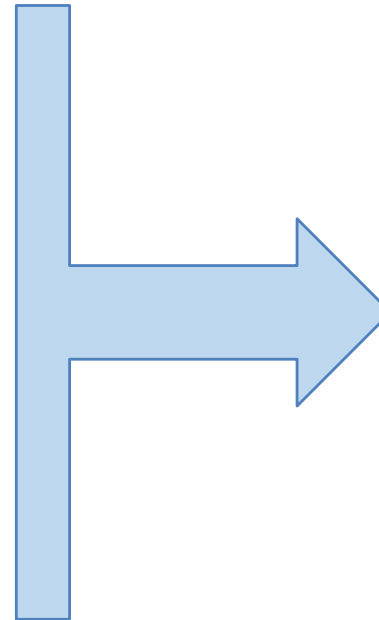
What color is the fire Hydrant on the left?



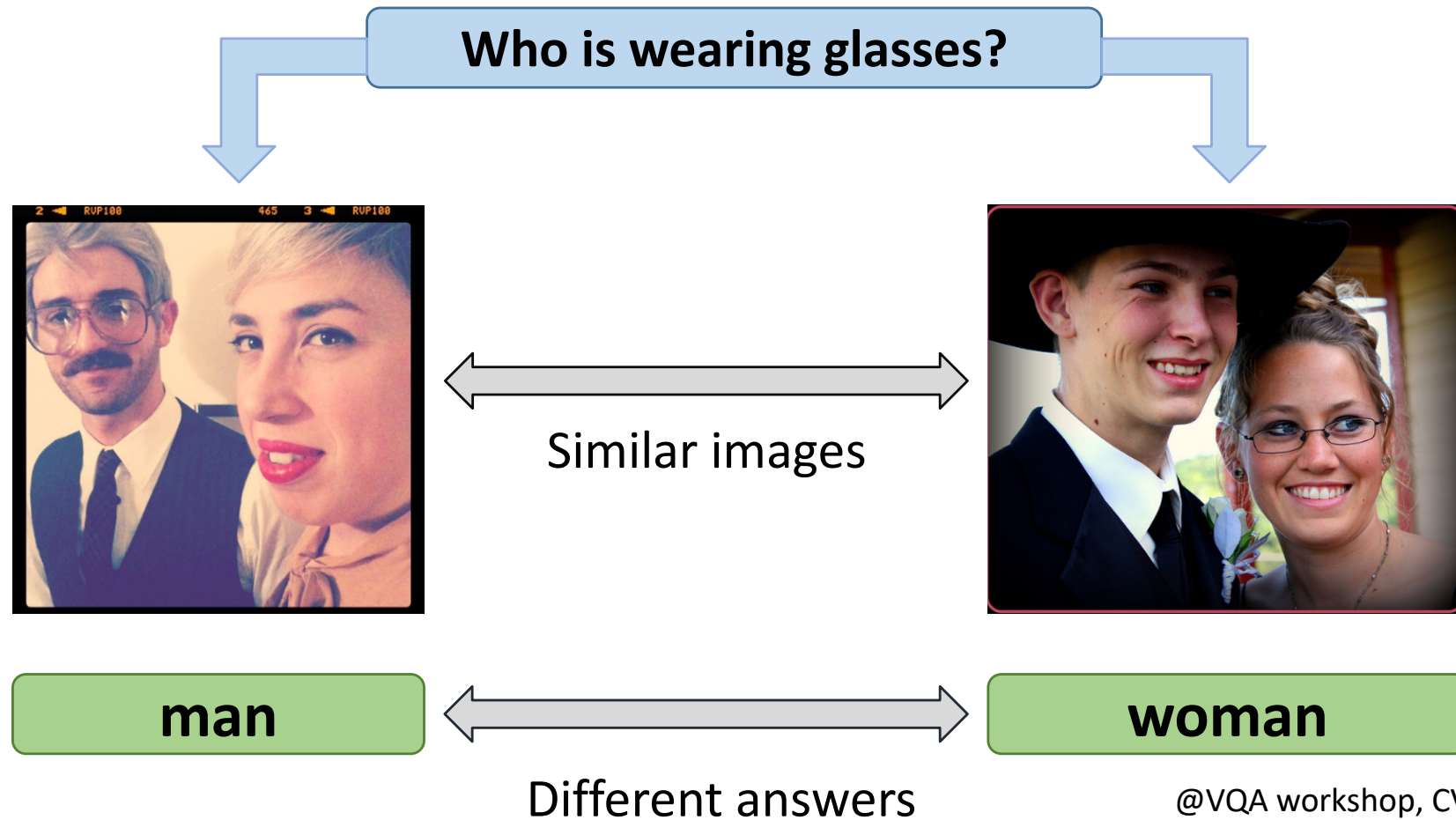
Green

VQA

What color is the fire Hydrant on the right?



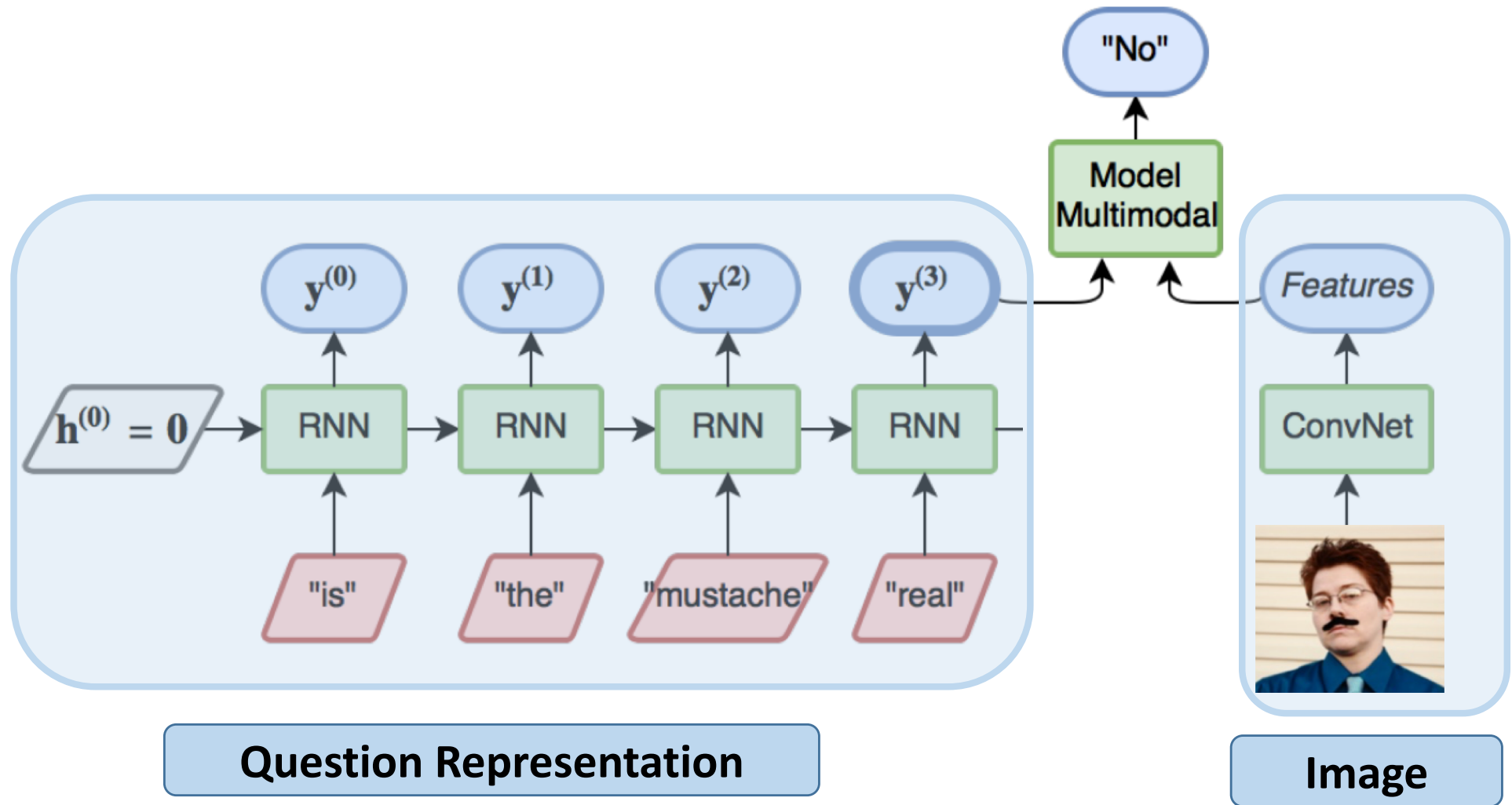
Yellow



@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

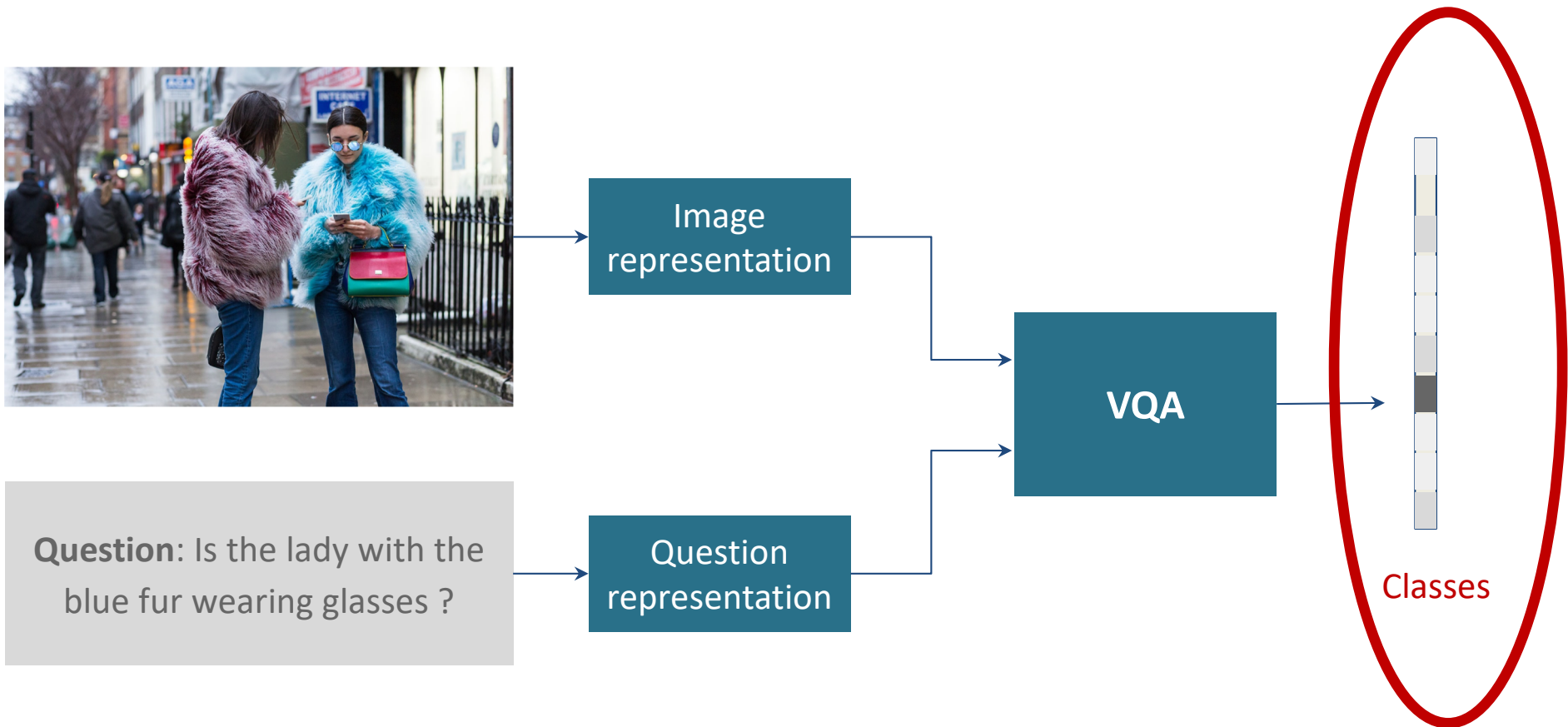


Question: Is the lady with the blue fur wearing glasses ?

VQA System

Yes

VQA: the output space



VQA processing

Image

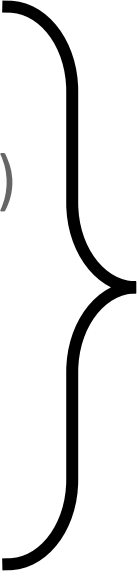
- Convolutional Network (VGG, ResNet,....)
- Detection system (EdgeBoxes, Faster-RCNN, ...)

Question

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

Learning

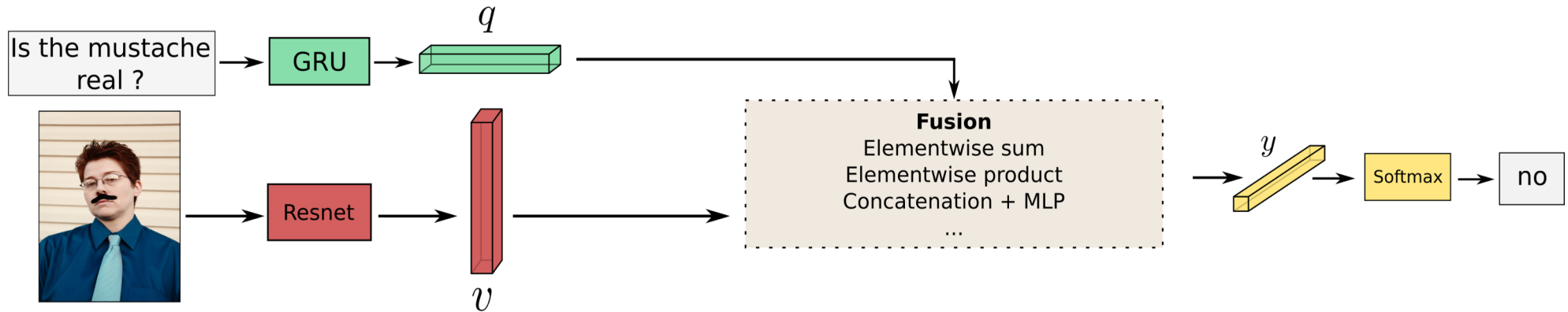
- Fixed answer vocabulary
- Classification (cross-entropy)



**Multimodal
Fusion
Reasoning**

Fusion in VQA

VQA: fusion



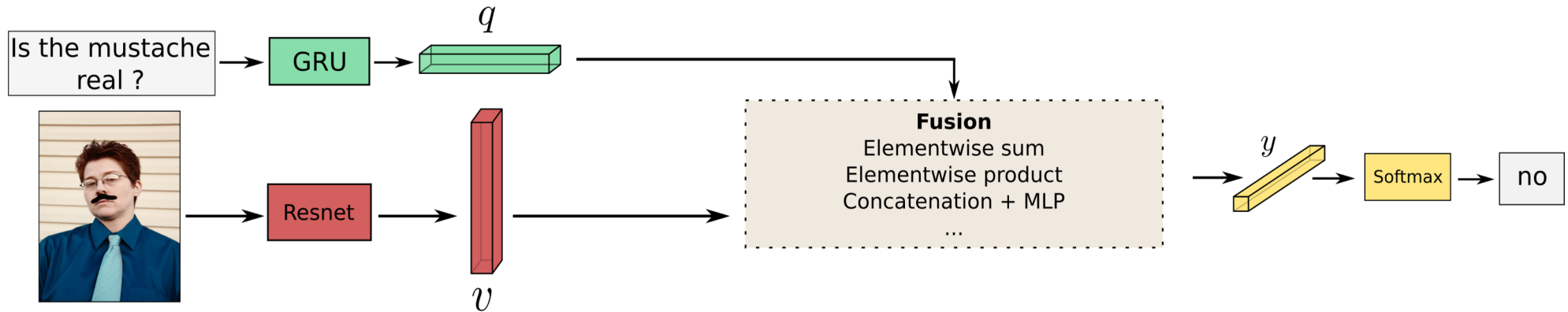
Concatenation & projection : $y = \mathbf{W} \begin{bmatrix} \mathbf{q} \\ \mathbf{v} \end{bmatrix}$

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

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

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

VQA: fusion



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

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

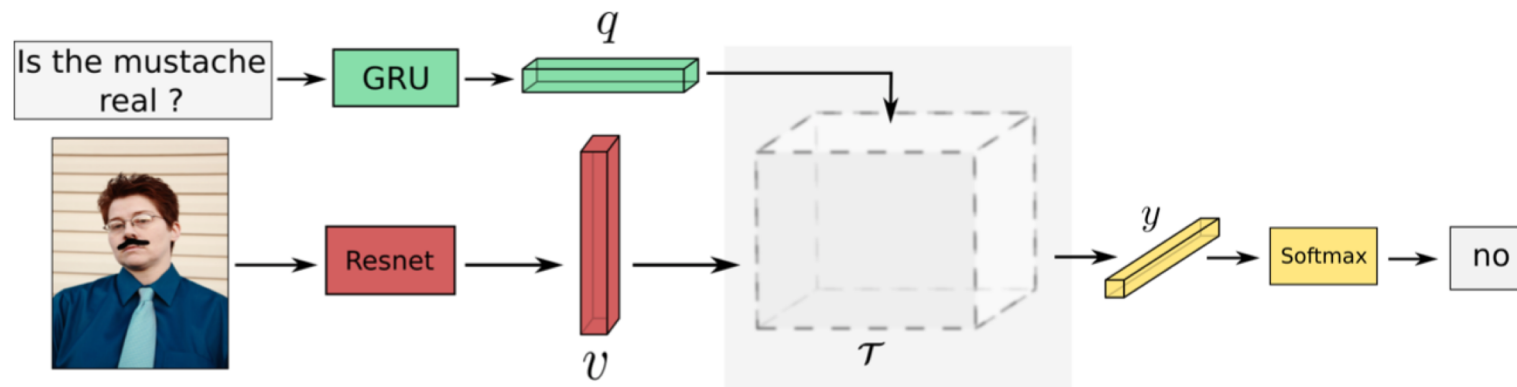
Element-wise product : $y = (\mathbf{W}\mathbf{q}) \odot (\mathbf{V}\mathbf{v})$

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

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:

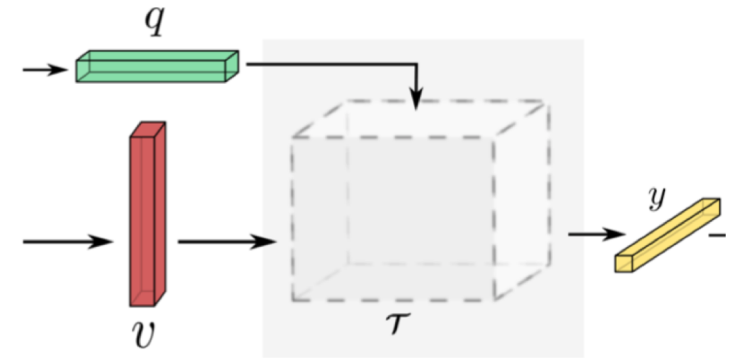
score for class k = bilinear combination of dimensions in q and v

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

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

VQA: bilinear fusion

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



Learn the 3-ways Tensor coeff.

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

Problem: \mathbf{q} , \mathbf{v} and \mathbf{y} are of dimension ~ 2000
 \Rightarrow **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

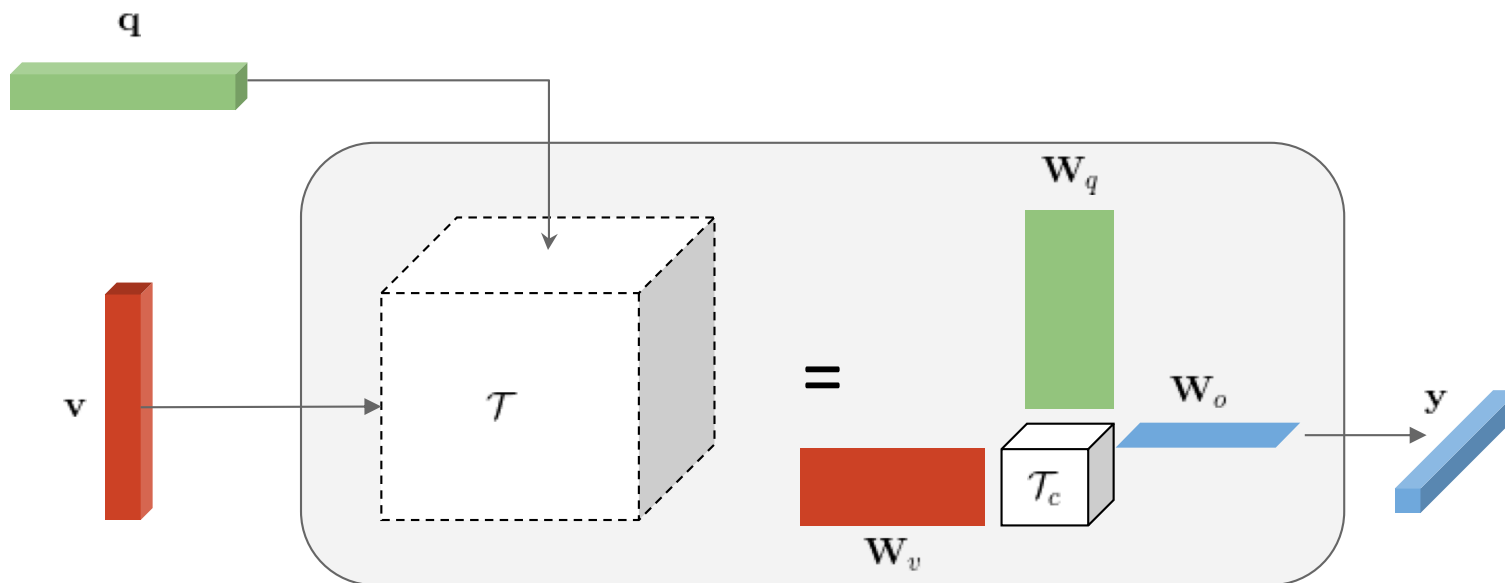
Tensor structure:

Tucker decomposition:

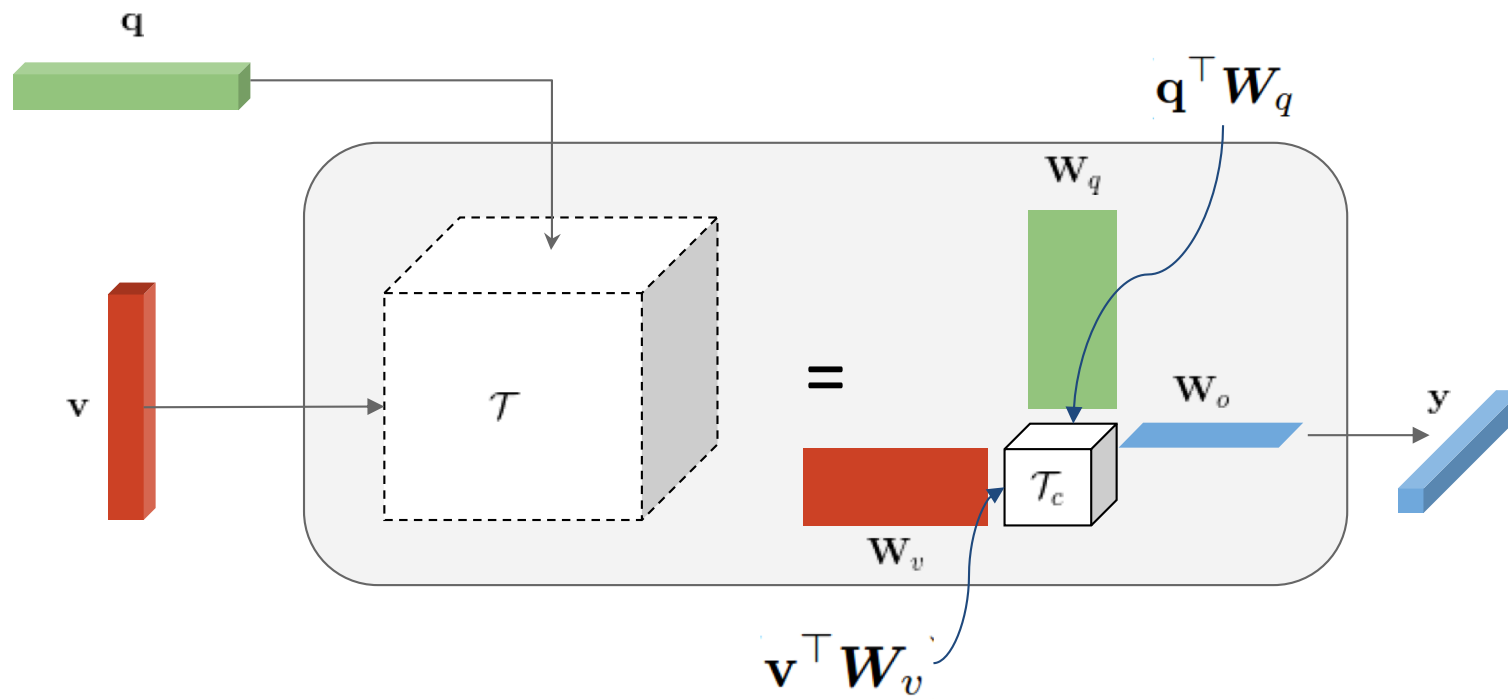
$$\mathcal{T} = ((\mathcal{T}_c \times_1 \mathbf{W}_q) \times_2 \mathbf{W}_v) \times_3 \mathbf{W}_o$$

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

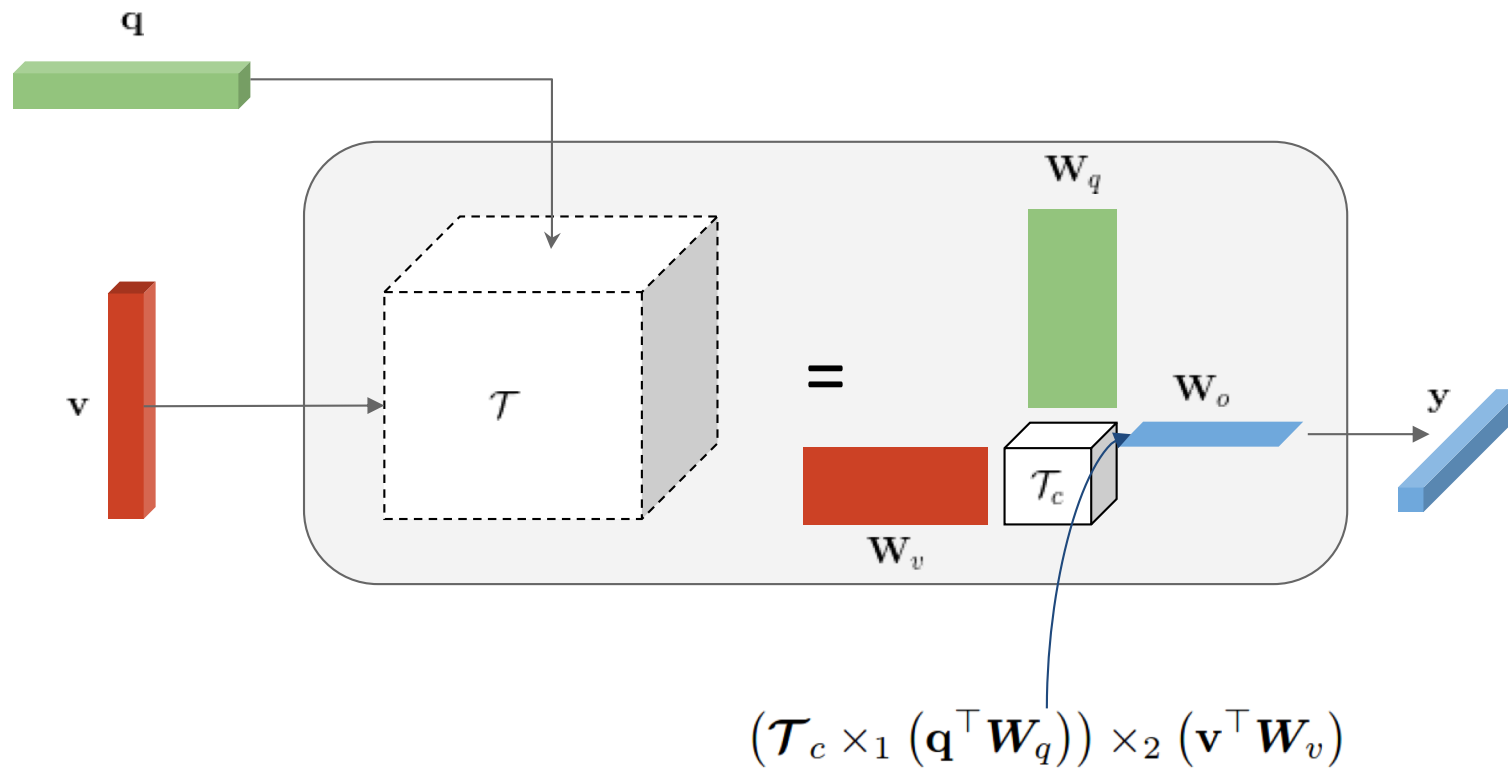
VQA: bilinear fusion



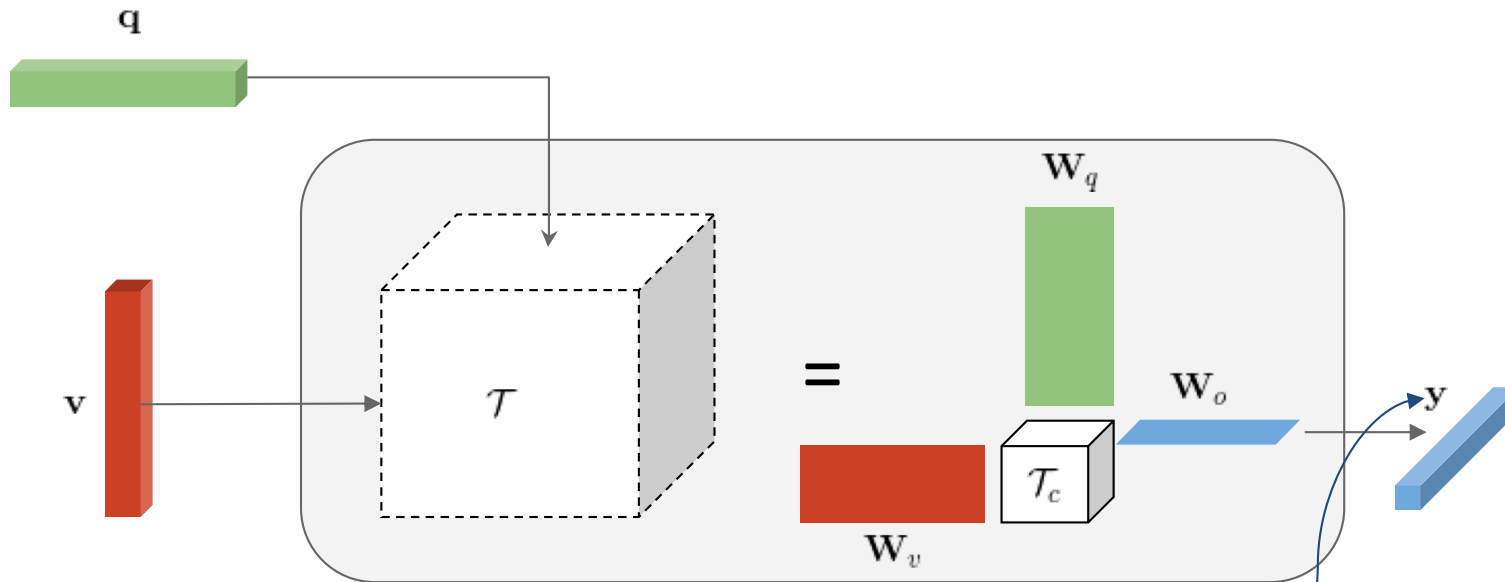
VQA: bilinear fusion



VQA: bilinear fusion



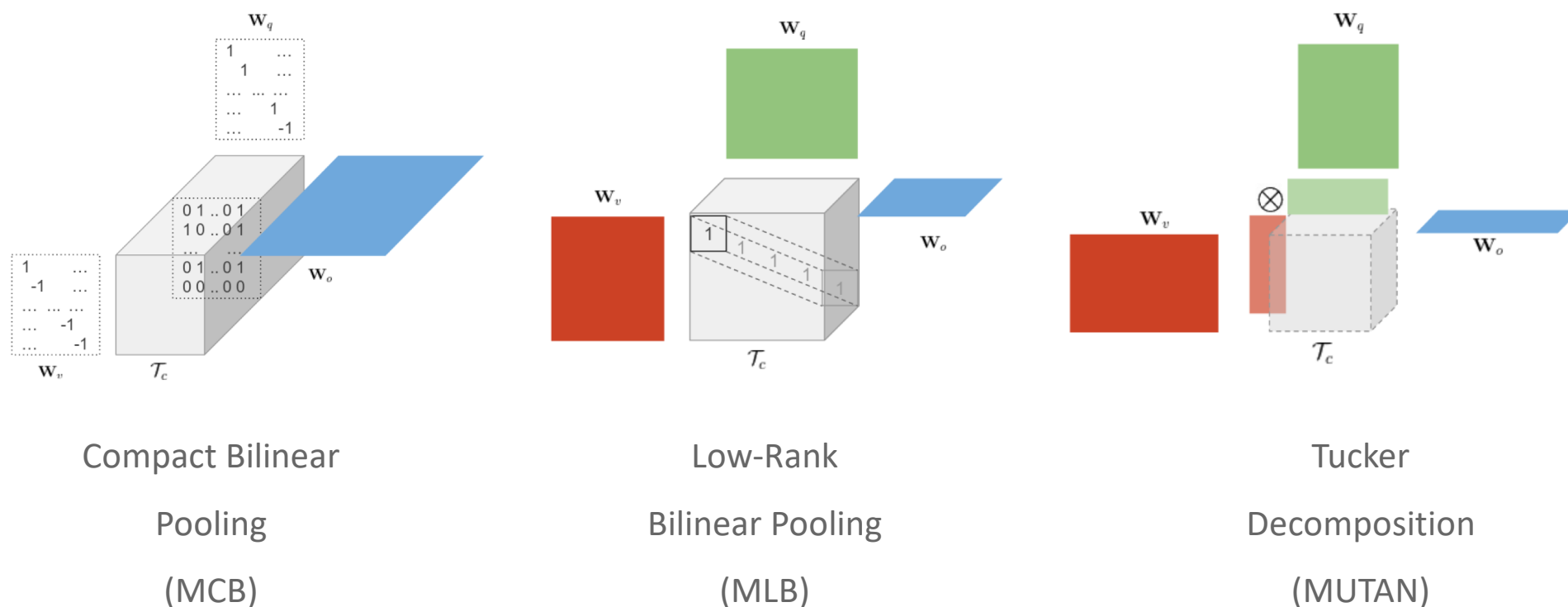
VQA: bilinear fusion



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

VQA: bilinear fusion

Other ways of structuring the tensor of parameters



VQA: bilinear fusion

MCB

Fukui, Park, Yang, Rohrbach et al, Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding, CVPR 2016

MLB

Kim et al., Hadamard product for low-rank Bilinear Pooling, ICLR 2017

MUTAN

Ben-younes H., Cadene et al., MUTAN: Multimodal Tucker Fusion for Visual Question Answering, ICCV 2017



MFB

Zhou et al., Multi-modal Factorized Bilinear Pooling with Co-Attention Learning for Visual Question Answering

MFH

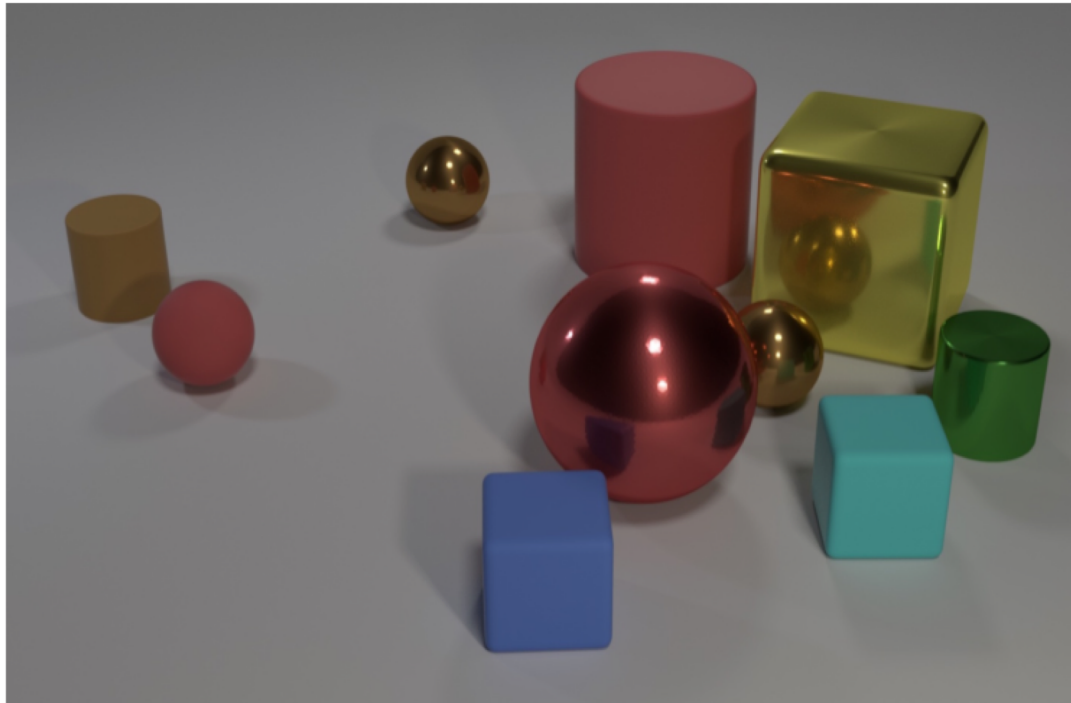
Zhou et al., Beyond Bilinear: Generalized Multi-modal Factorized High-order Pooling for Visual Question Answering

VQA: bilinear fusion

	$ \Theta $	All
Sum	8M	58.48
MCB* [10]	32M	61.23
Concat MLP ¹	13M	63.85
Tucker* [4]	14M	64.21
MLB* [16]	16M	64.88
MUTAN* [4]	14M	65.19
MFB* [30]	24M	65.56
MFH* [32]	48M	65.72

Comparing fusion schemes on the VQA2.0
Dataset

Reasoning in VQA



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

VQA: reasoning

What is reasoning (for VQA)?

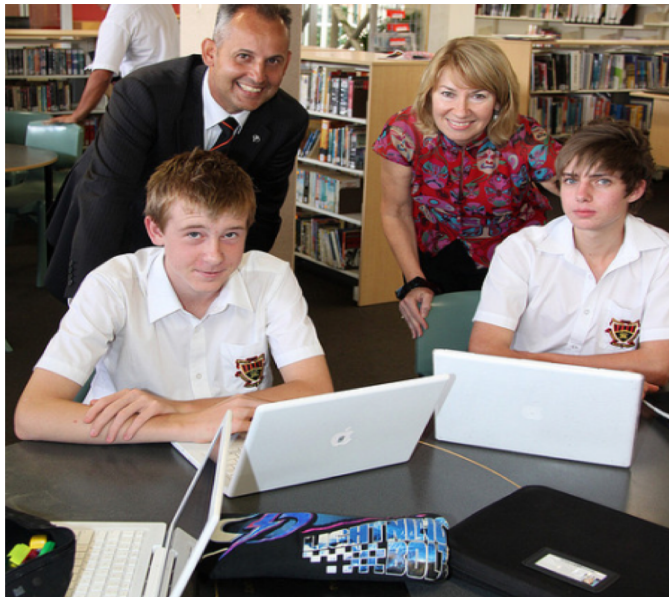
- **Attentional reasoning:** given a certain context (i.e. Q), focus only on the relevant subparts of the image
- **Relational reasoning:** object detection + mutual relationships (spatial, semantic,...), merging both with Q
- **Iterative reasoning:** refining the attention step-by-step, each time extracting a different piece of information from the image
- **Compositional reasoning:** use Q to 1/ select elementary attentional blocks, and 2/ assemble their predictions

VQA: attentional reasoning

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

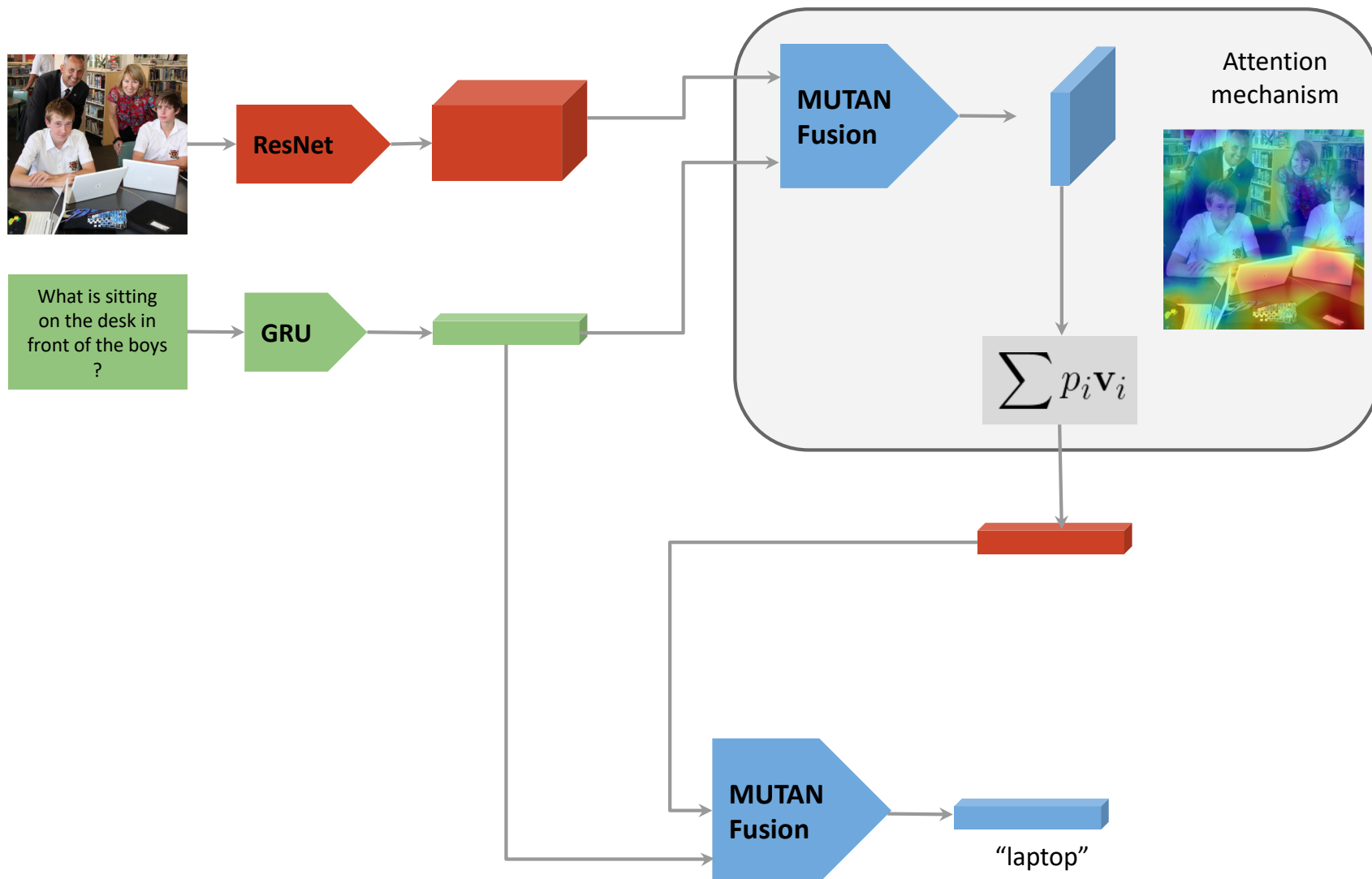
- Each region scored according to the question

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



- Representation = sum of all (weighted) embeddings

VQA: attentional reasoning

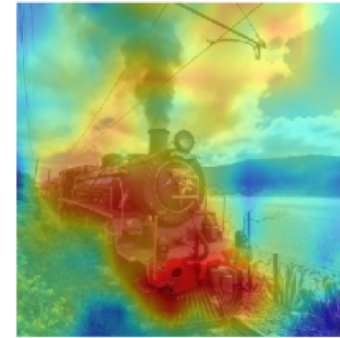
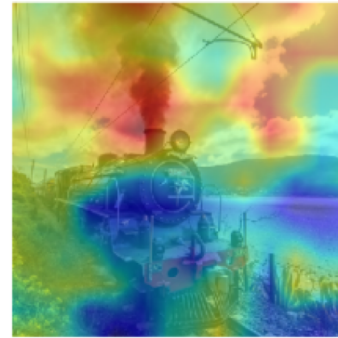


Ben-younes H.* Cadene R.*, Thome N., Cord M., *MUTAN: Multimodal Tucker Fusion for Visual Question Answering*, ICCV 2017

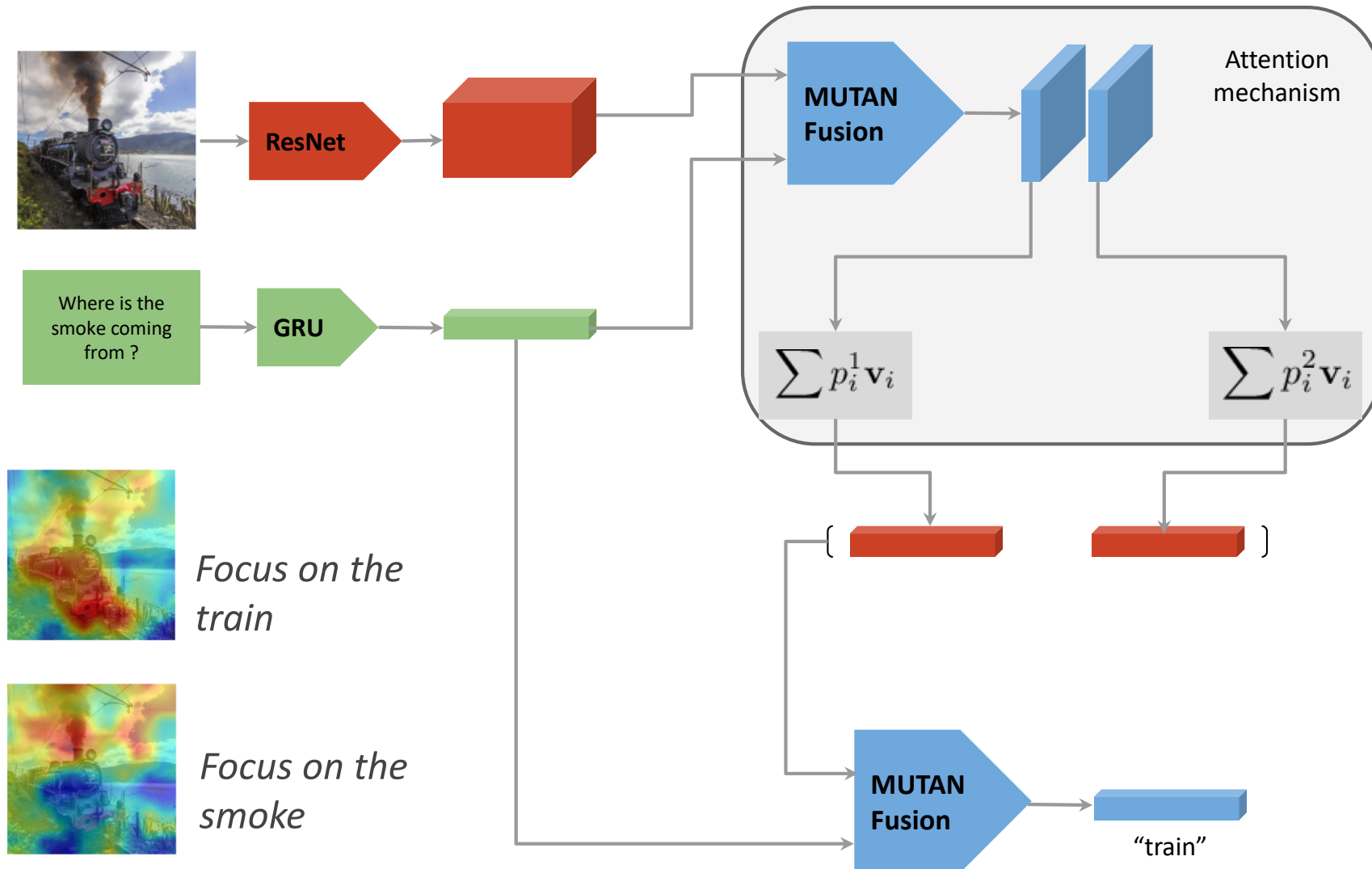
Multi-glimpse attention

We might want to focus separately on multiple regions

Where is the smoke
coming from ?



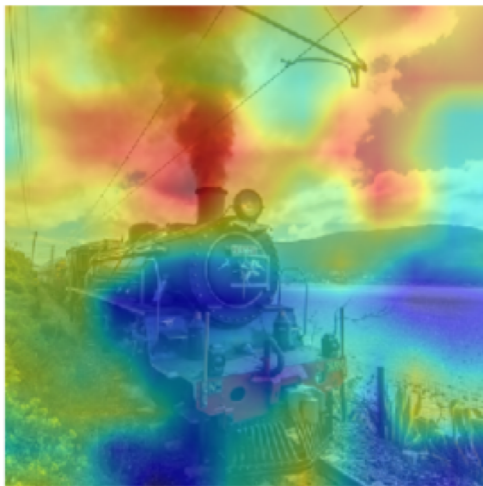
Multi-glimpse attention



VQA 2.0 - Some examples

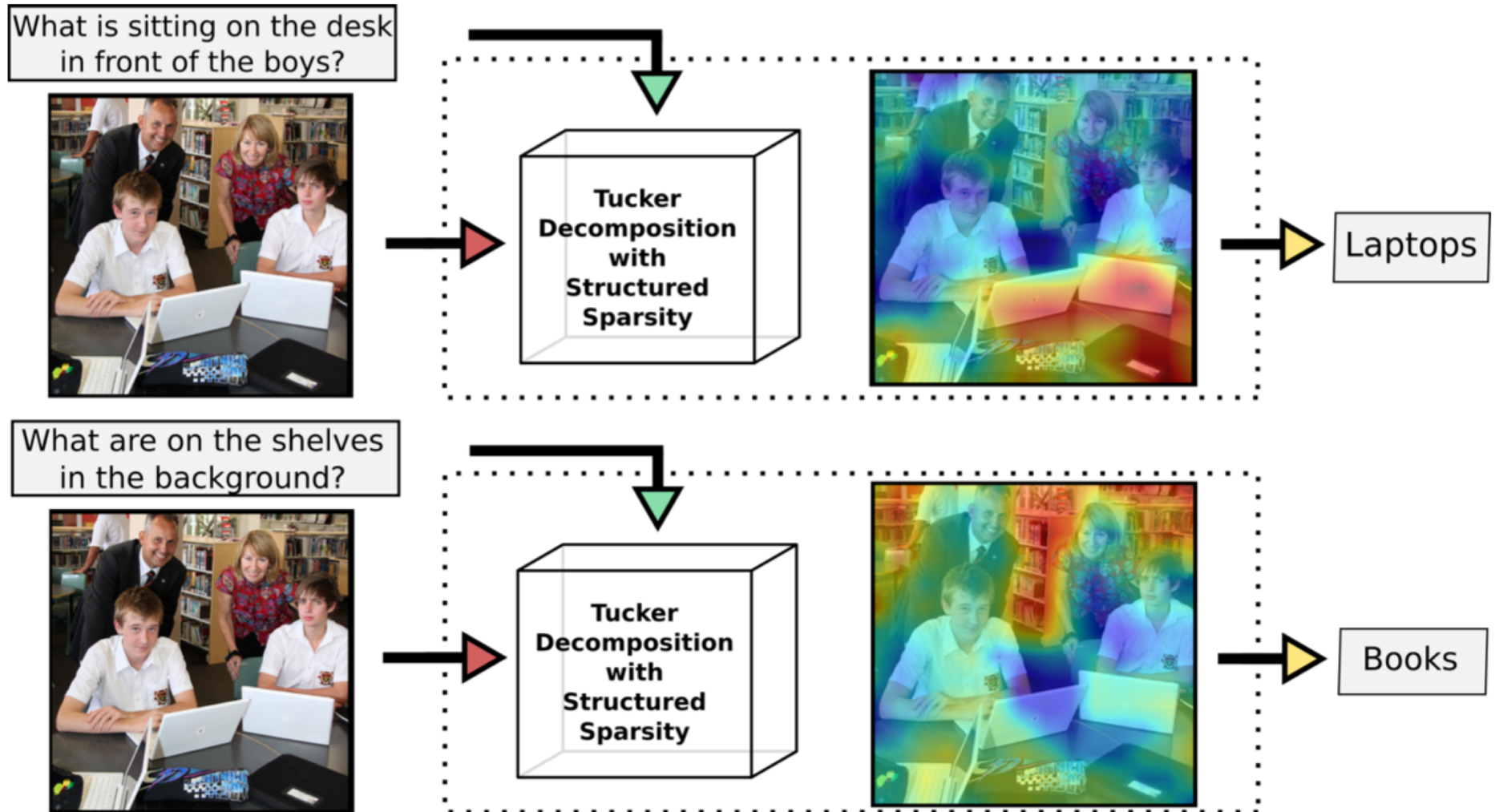


(a) Question: Where is the woman ? - Answer: on the elephant



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

VQA: Attention process & reasoning



VQA: Attention process & reasoning

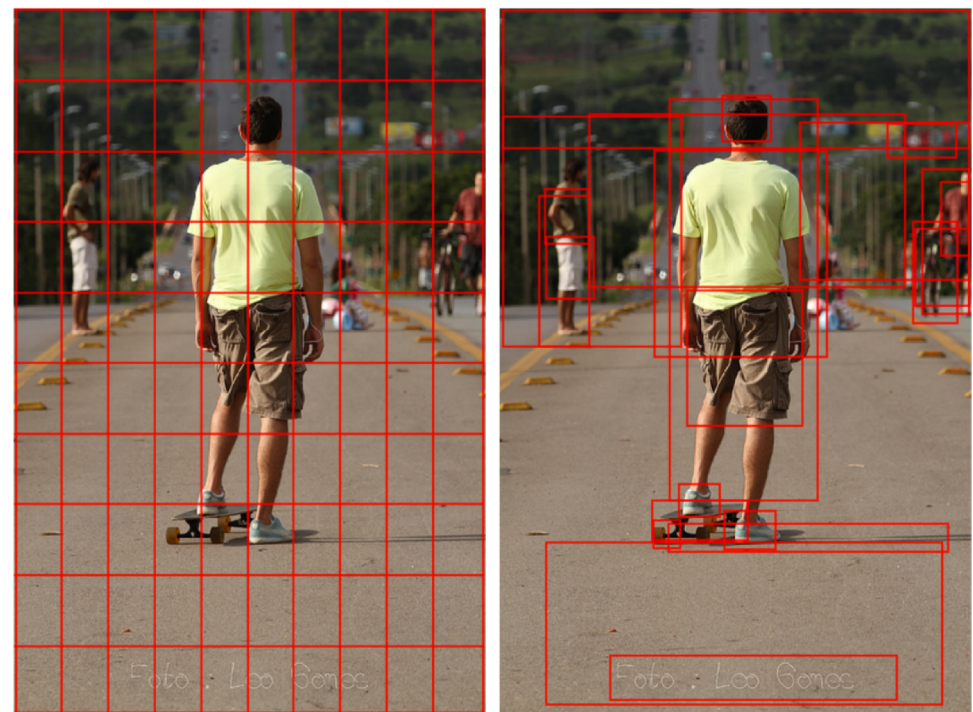
Evaluation on VQA dataset:
Best MUTAN score of
67.36% on test-std

Human performances about
83% on this dataset

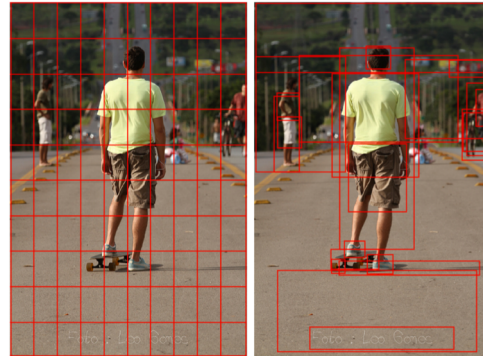
The winner of the VQA
Challenge in CVPR 2017
(and CVPR 2018) integrates
adaptive grid selection from
additional region detection
learning process

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

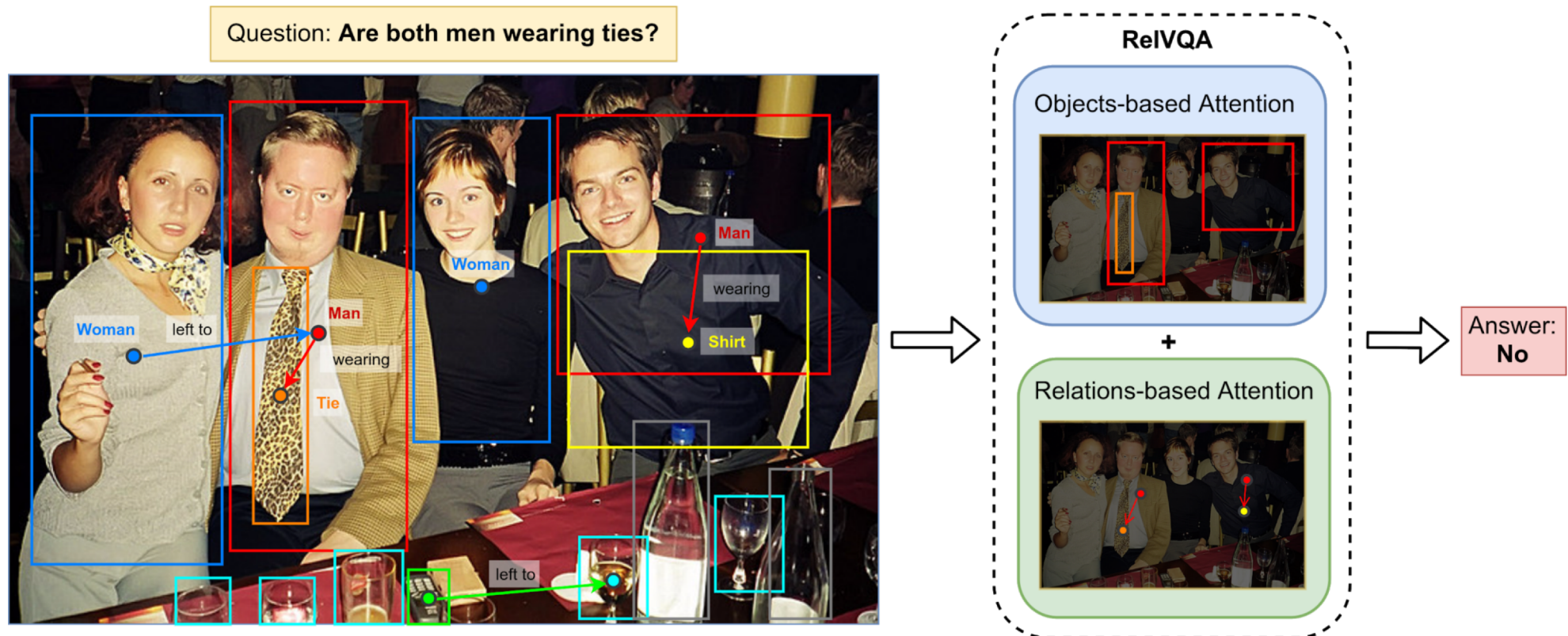
Peter Anderson^{1*}, Xiaodong He², Chris Buehler², Damien Teney³
Mark Johnson⁴, Stephen Gould¹, Lei Zhang²



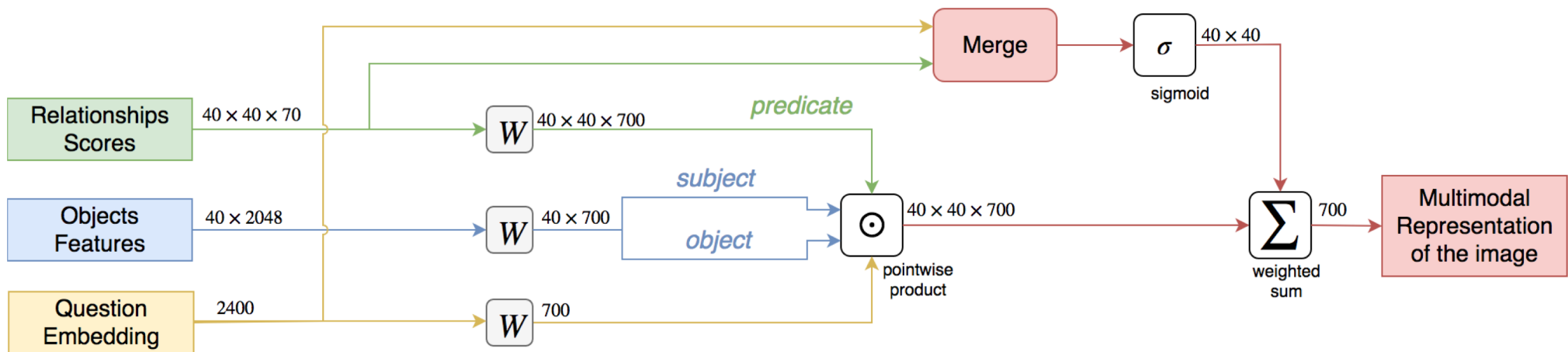
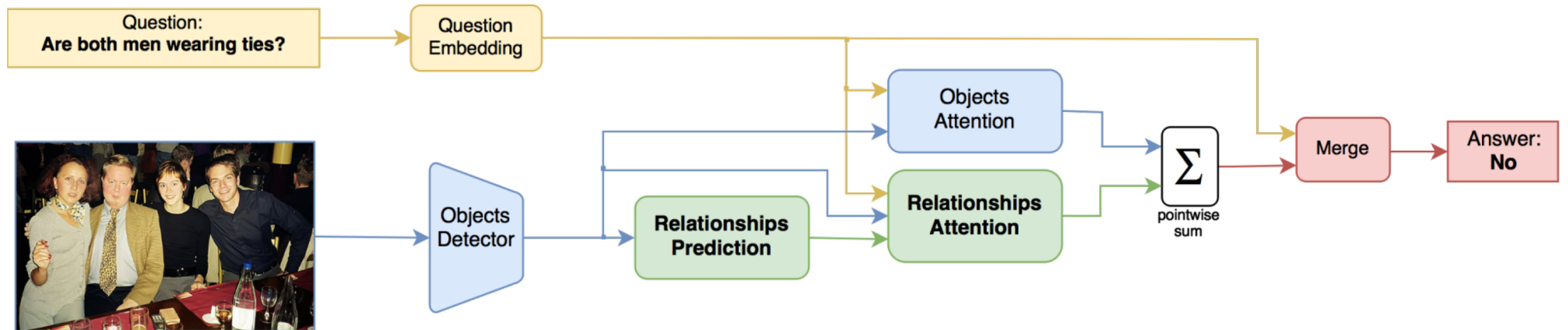
Bottom-up and Relational reasoning



Determine the answer using relevant objects and relationships



Bottom-up and Relational reasoning



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

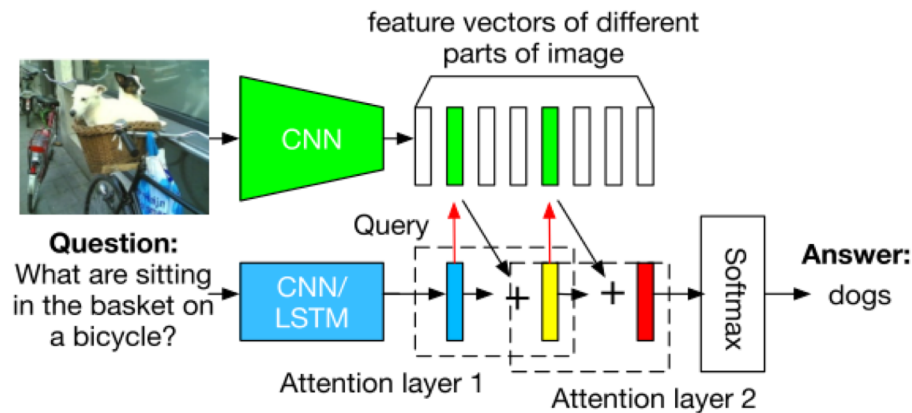


What are sitting in
the basket on a
bicycle ?

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

Stacked Attention

At each step, the query-attention process extracts more fine-grained visual information



(a) Stacked Attention Network for Image QA



Original Image First Attention Layer Second Attention Layer

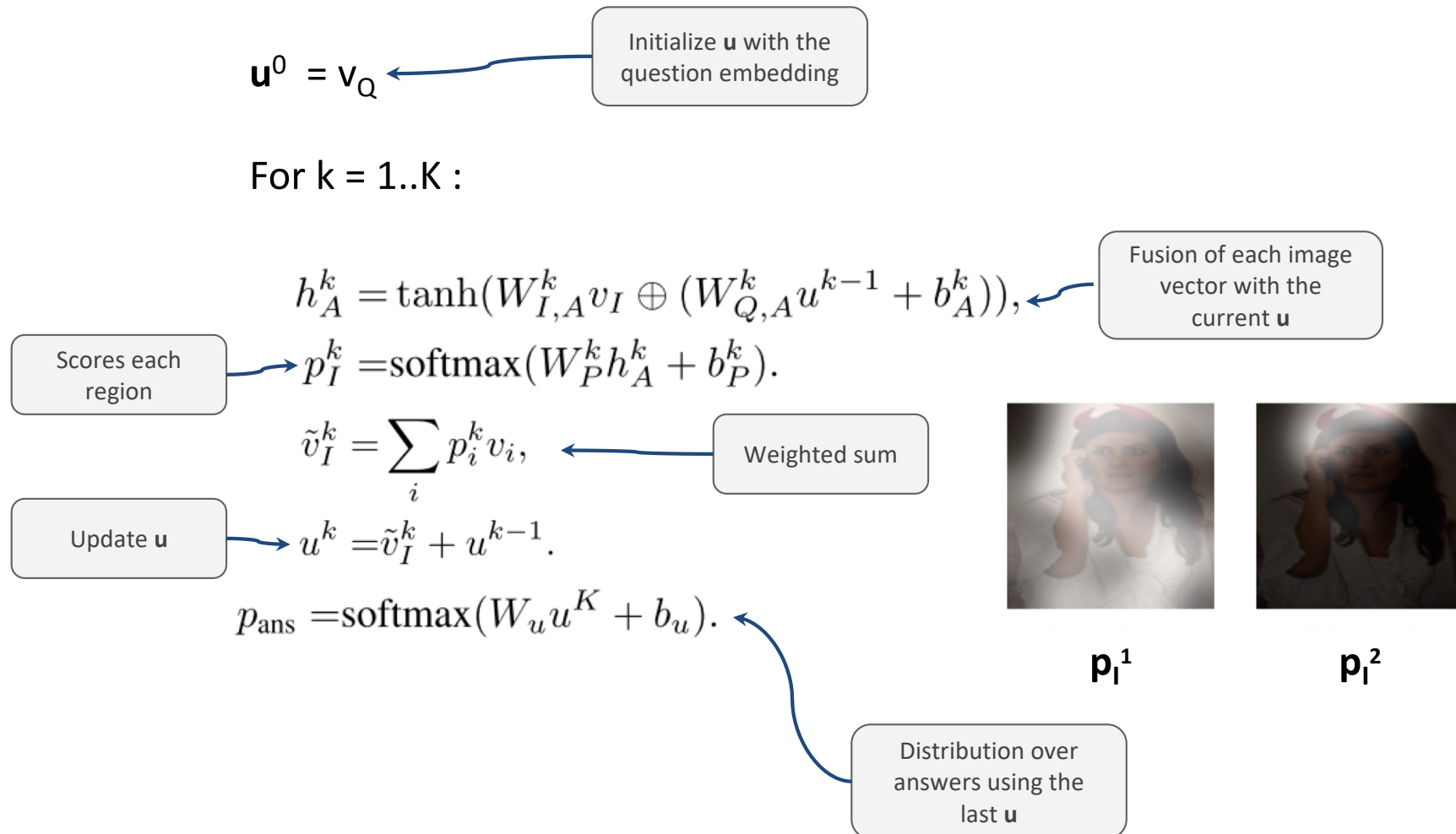
(f) What is the color of the horns?
Answer: red Prediction: red



(b) What is the color of the box ?
Answer: red Prediction: red



Stacked Attention



CLEVR Dataset

Question: *What covers the ground ?*

VQA System:

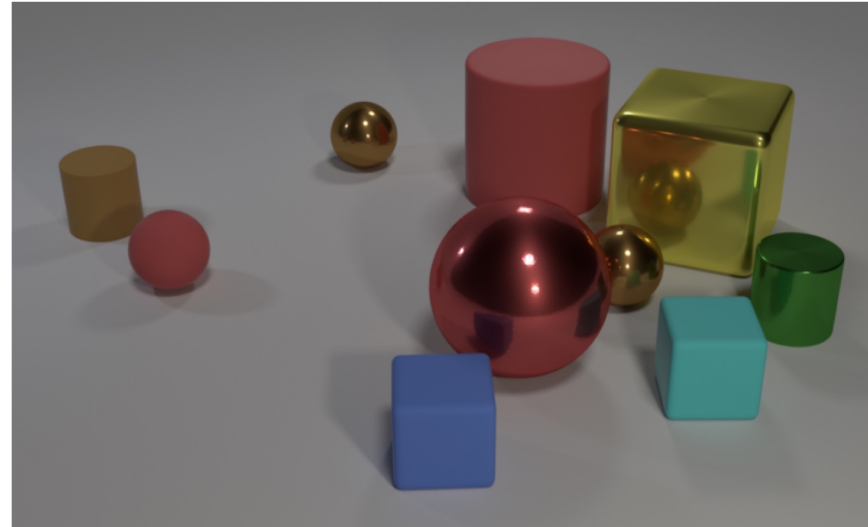
“I don’t even need to look at the image, let’s just answer *snow*, as usual”

Johnson et al., CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning

CLEVR Dataset

The model need to be able to:

- Count
- Extract attributes
- Compare
- Perform logical operations
- Use memory



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

Q: **What size** is the **cylinder** that is **left of** the **brown metal** thing that is **left of** the **big sphere**? Q: There is a **sphere** with the **same size** as the

metal cube; is it **made of the same material as** the **small red sphere**?

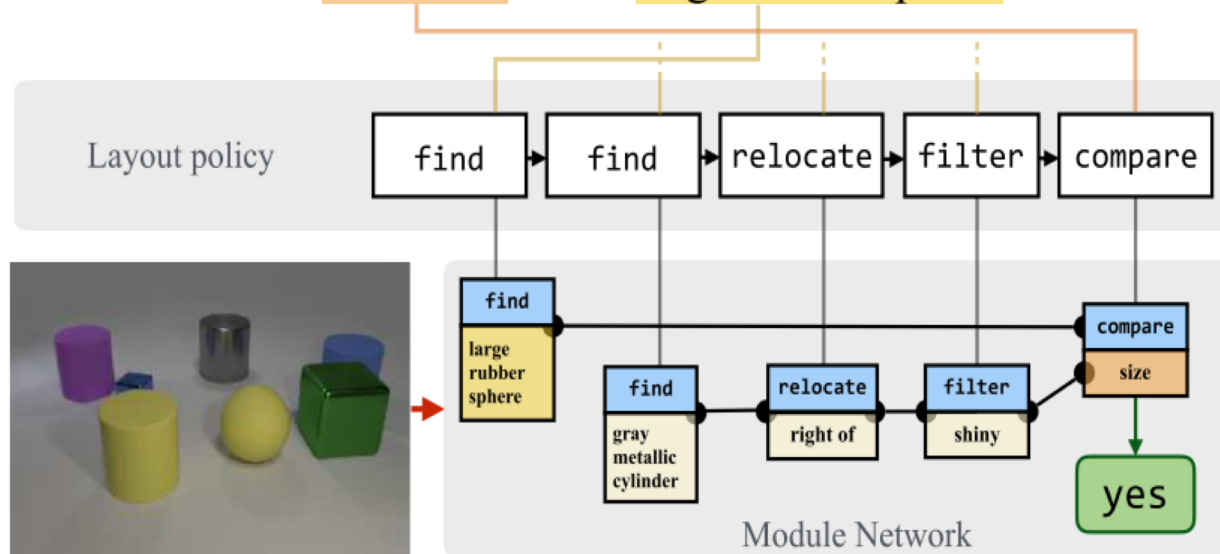
Q: **How many** objects **are either** **small cylinders** **or** **metal things**?

Compositional reasoning

Decomposing Q into multiple elementary operations

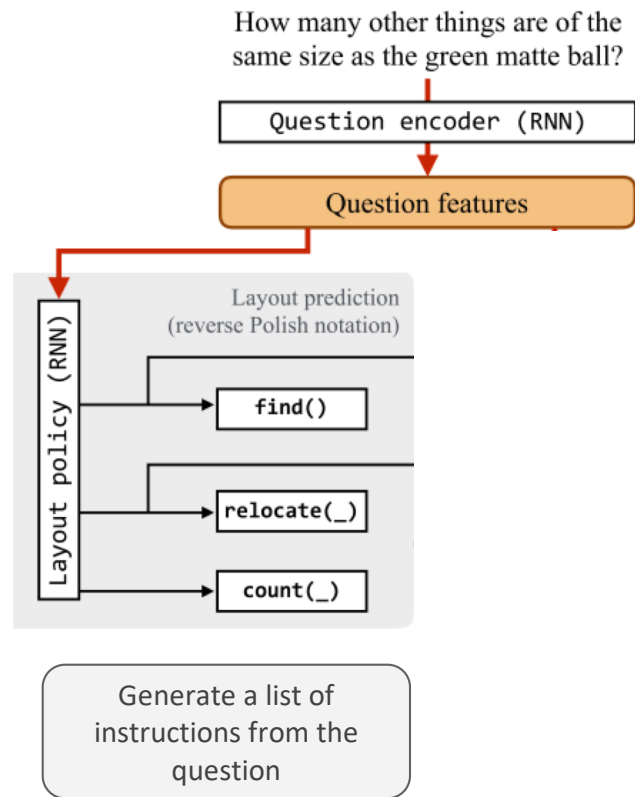
- Each operation corresponds to a visual module
- Jointly learn each module's weight **and** how to assemble the modules given a question

There is a shiny object that is right of the gray metallic cylinder; does it have the same size as the large rubber sphere?

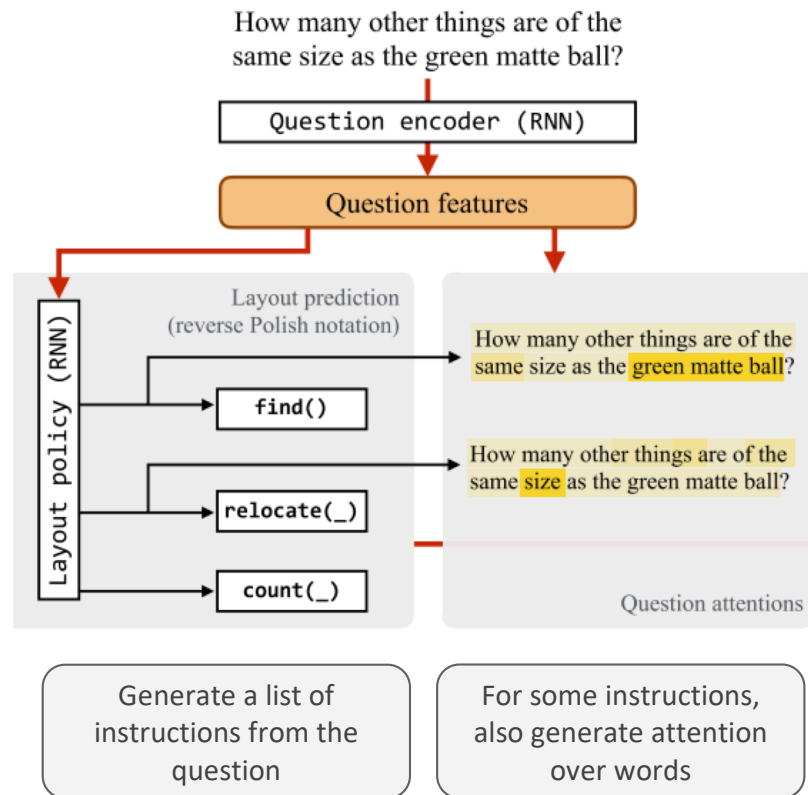


Ronghang Hu et. al., *Learning to Reason: End-to-End Module Networks for Visual Question Answering*, ICCV 2017

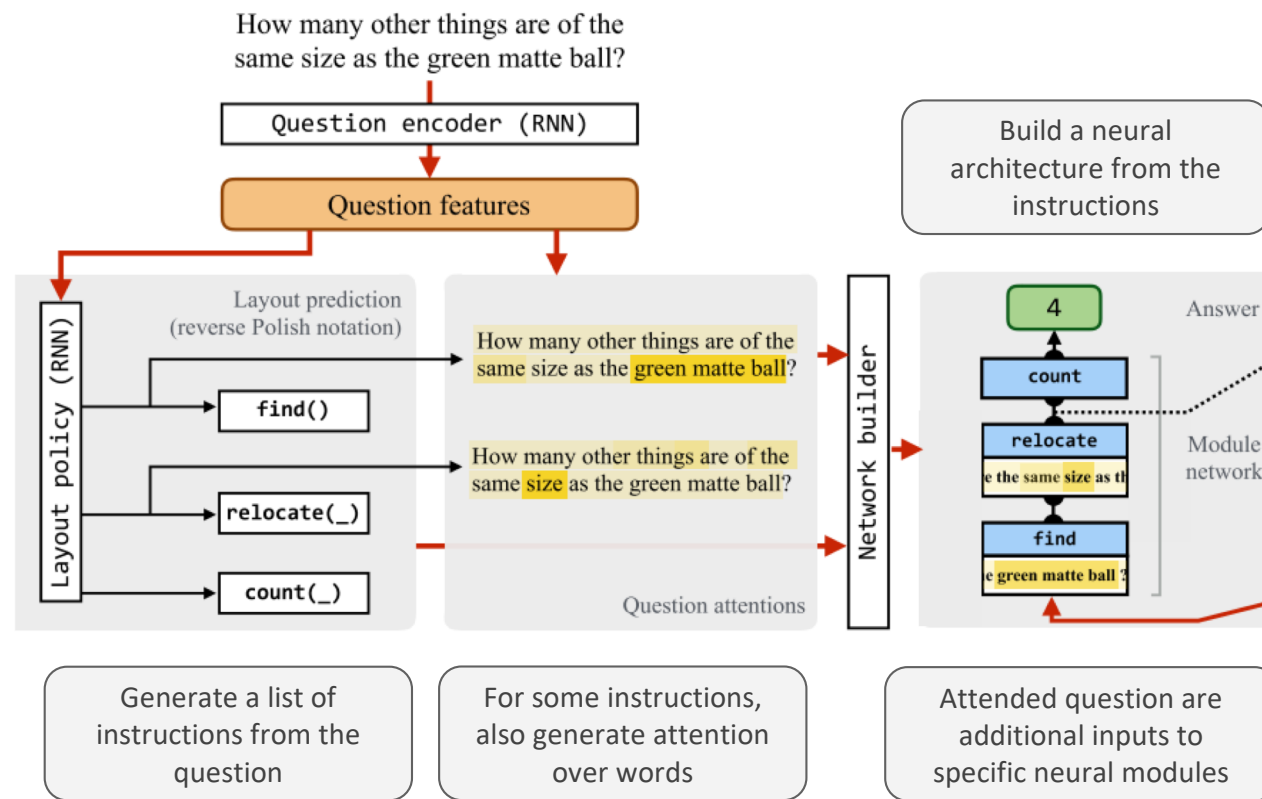
Compositional reasoning



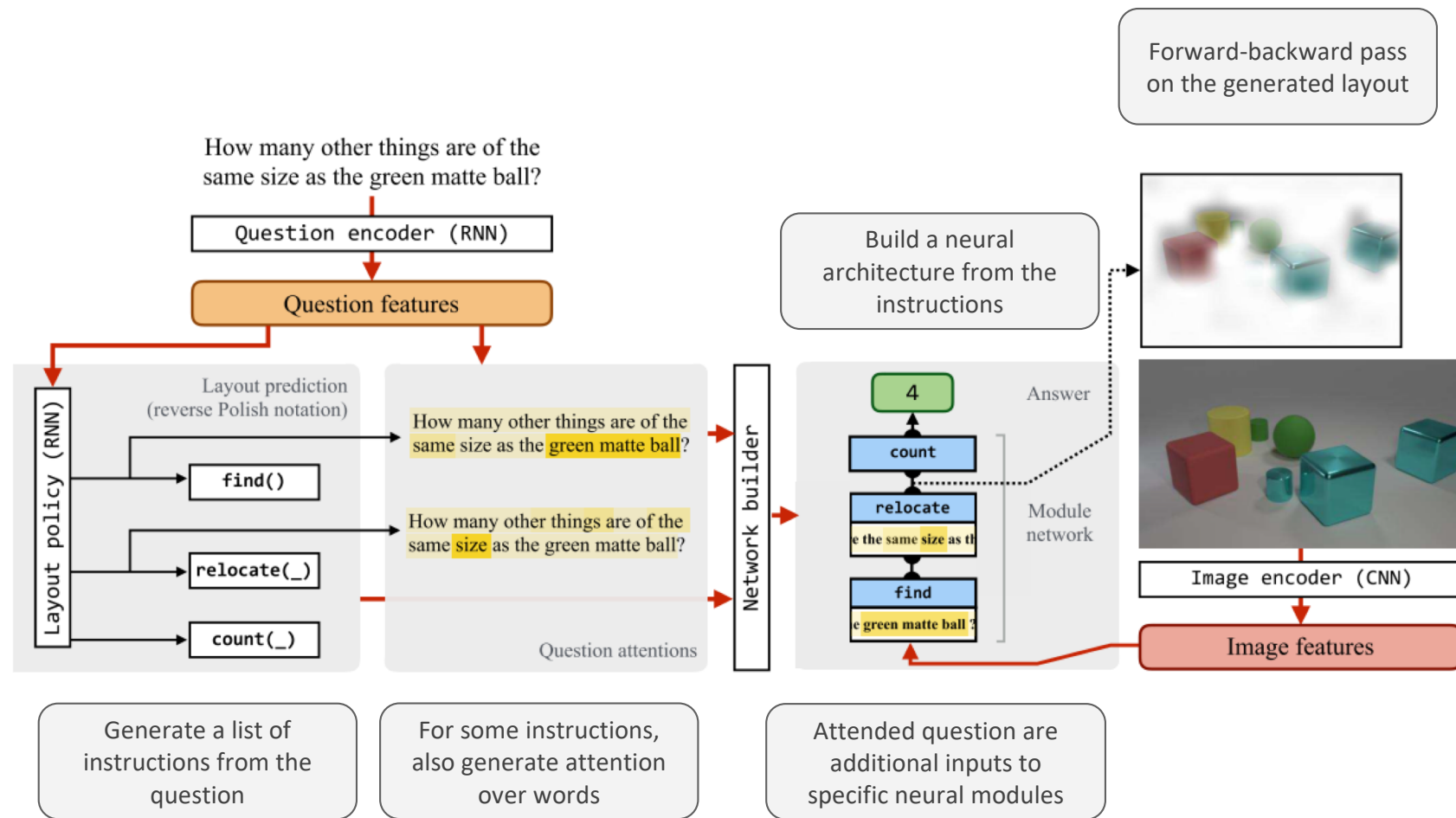
Compositional reasoning



Compositional reasoning



Compositional reasoning



Conclusion

Complex models mixing:

- Unimodal representations
 - CNN, Faster-RCNN,...
 - RNN, LSTM, GRU, SRU ...
- Multimodal fusion
 - Linear, Deep, Bilinear, ...
- Reasoning
 - Attention: Simple, Multi-glimpse, Stacked,...
 - Relational
 - Compositional: program generation

Complex datasets

- Bias in the annotation
- Non-trivial evaluation
 - How can we automatically say that an answer is false ?

VQA: Attention process & reasoning

Many initiatives to improve datasets and evaluate reasoning as:

VQA v2.0 dataset and challenge 2017

[Y. Goyal, D. Batra, D. Parikh, CVPR 2017]

CLEVR dataset [J. Johnson et al, CVPR 2017]

- Questions about visual reasoning including attribute identification, counting, comparison, spatial relationships, and logical operations.

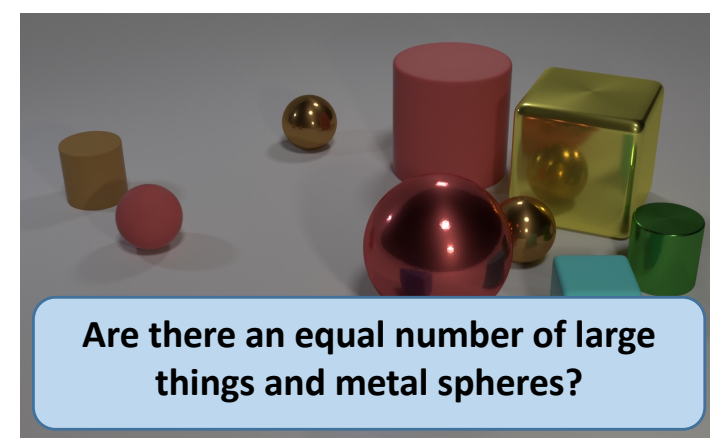
TDIUC dataset and challenge (Task Driven Image Understanding Challenge)

- Over 1.6 million questions organized into 12 different categories

Visual dialogue task: a novel task that requires an AI agent to hold a dialog with humans in natural, conversational language about visual content.



Figure 1: Examples from our balanced VQA dataset.



MLIA/Chordettes team: Matthieu Cord <http://webia.lip6.fr/~cord>

D.Picard (CNRS deleg), N. Thome (associate member), Arnaud Dapogny (Postdoc)

PhD T. Robert, T. Mordan, M. Blot, M. Carvahlo, H. BenYounes, R. Cadene, E. Mehr, M. Engilberge,
Y. Chen, A. Saporta (ing)

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

