Learning Multimodal Word Representations: Language Grounding in Visual Context and Visual Question Answering

12-12-2017

Patrick Gallinari

Université Pierre et Marie Curie – Paris 6, France

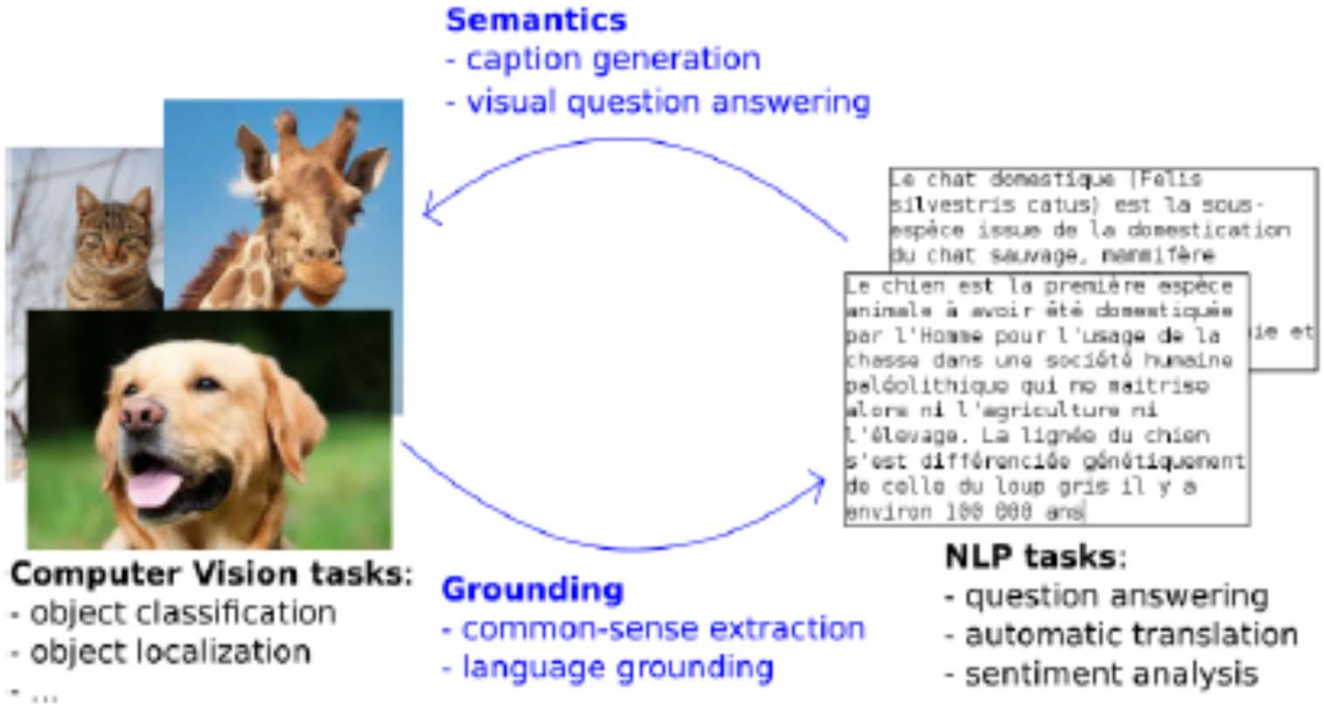
patrick.gallinari@lip6.fr

UPMC SORBONNE UNIVERSITÉS

MLIA Machine Learning & Deep Learning for Information Access

General context

- Multimodal machine learning with text and image sources
  - Complementary role of text and image
  - Improve language/ image processing tasks



General context

- Distributional hypothesis
  - Text
    - Linguistic items with similar **context** distributions should have similar **meanings**
    - Meaning is represented by a vector summarizing the context distribution → semantic space
      - Similarity measures in the semantic space represents word meaning similarity
    - Meaning is entirely defined by co-occurrence patterns and not by real world situations – similar to AI symbol grounding problem
    - Bruni et al 2012 -> according to semantic similarity sky is green, flour is black, etc
  - Image
    - Object recognition methods extract vector based representations from natural images
      - Bag of visual words (image patches)
      - Convolutional Neural Networks introduce hierachical visual vectors representations of images
    - Distributional hypothesis for images (e.g. Bruni et al 2012)
      - Semantically similar objects tend to occur in similar environments in images
      - Importance of the visual context

General context: examples

- Grounded Language Model
  - Caption generation: text generation learned from image representations
  - Visual QA: classification task, classifiers learned from joint image + query text representation

Fig. Vinyals et al. 05

Figure 5. A selection of evaluation results, grouped by human rating.


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 Real

Visual Genome

General context: examples

- Visual Common Sense
  - Modeling text-image bias
    - Frequency with which one refers to things or actions does not correspond to real word frequencies.
    - People tend to eliminate common things
    - Introduces a bias of language wrt the « real world frequencies »
    - Both textual and image data have biases about the information they encode about context
      - Motivates multi-modal approaches for learning word representations

Fig. from Misra et al. 2016
Illustrating human reporting bias for image annotation

General context: example

- Cross modal mapping (here image to text)
  - Objective
    - Learn a textual description of an image
      - i.e. using an image as input, generate a sentence that describes the objects and their relation!
  - Neural image caption generator (Vinyals et al. 2015)
    - Inspired by a translation approach but the input is an image
      - Use a RNN to generate the textual description, word by word, provided a learned description of an image via a deep CNN
    - Makes use of attention mechanism
    -

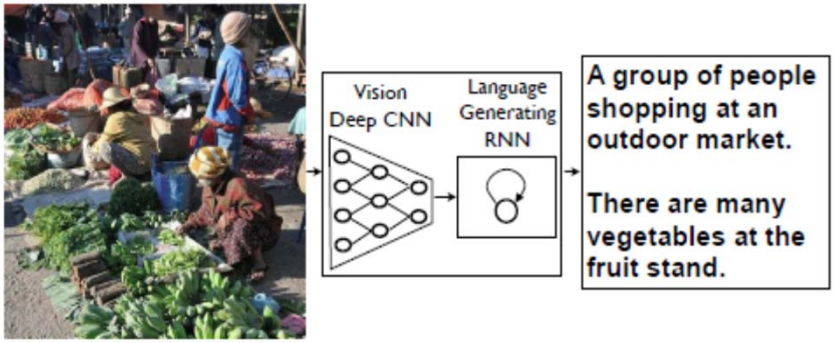
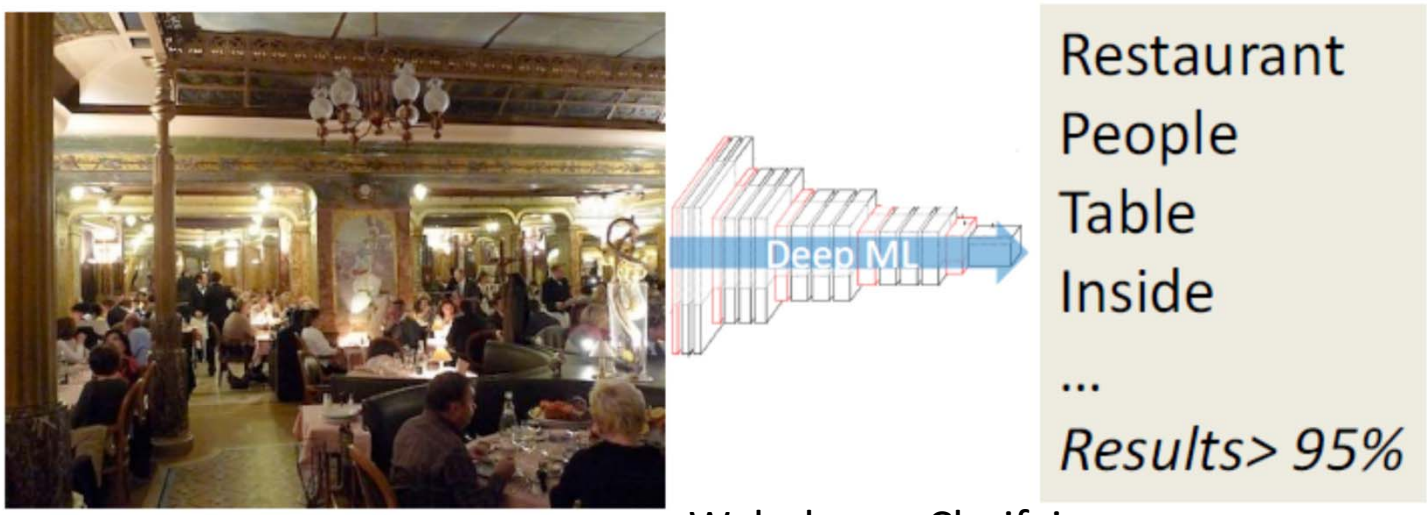


Figure from Vinyals et al. 2015

Figure 1. NIC, our model, is based end-to-end on a neural network consisting of a vision CNN followed by a language generating RNN. It generates complete sentences in natural language from an input image, as shown on the example above.

Background- Models – CNNs for classification and segmentation

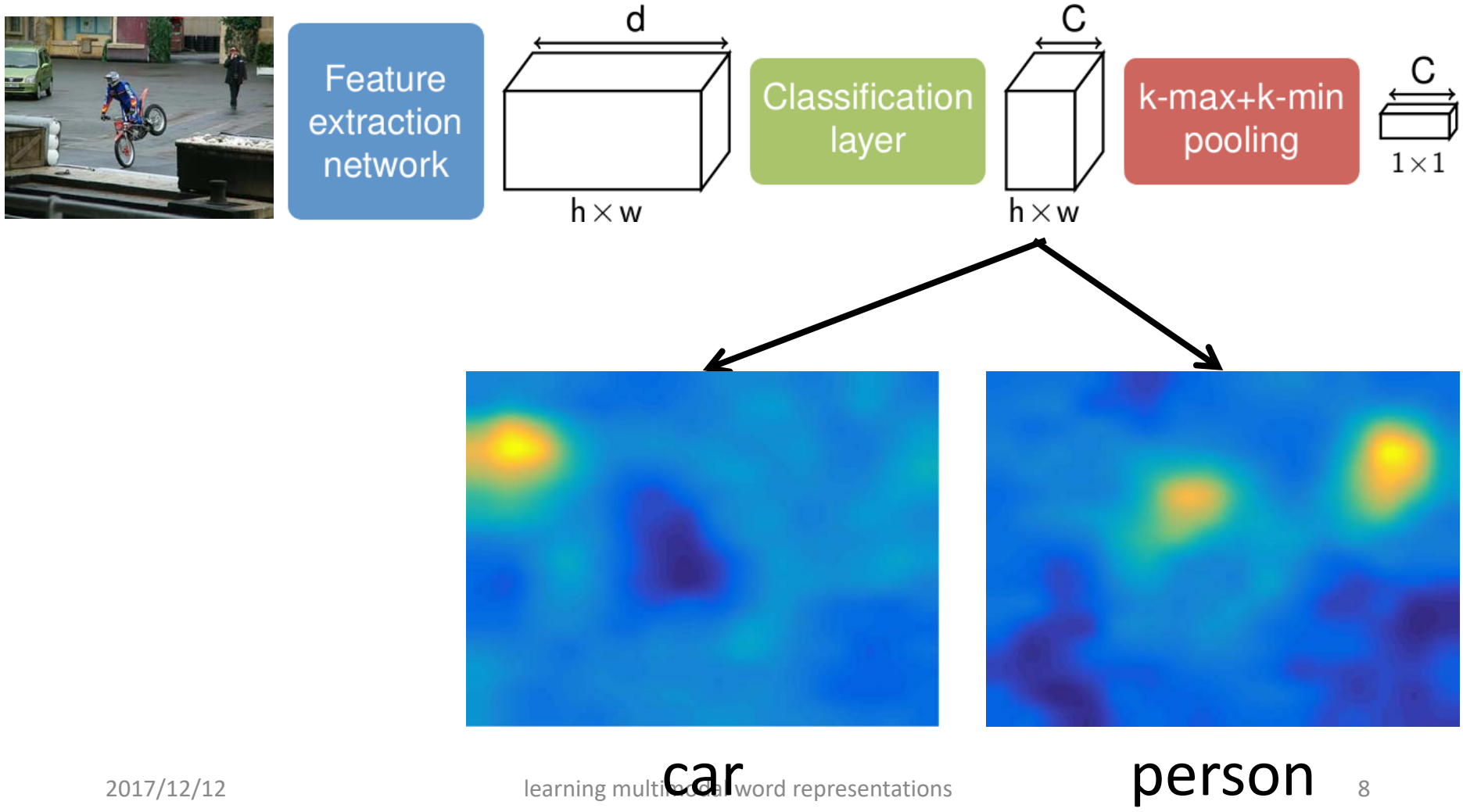
- CNNs state of the art models for image labeling and segmentation



Web demo: Clarifai

- Layers build increasingly abstract vector representations of input image
- Learned representations on large datasets (Imagenet) may be used for other tasks or other datasets

Background- - Models – CNNs for classification and segmentation (Durand 2016)



Background- - Recurrent neural networks Language models

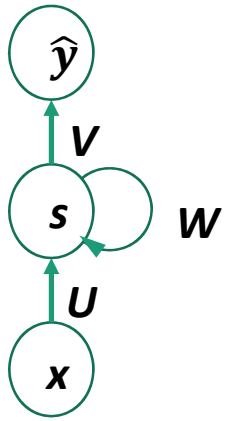
- Objective:
  - Probability models of sequences $(x^1, x^2, \dots, x^t)$
  - Items may be words or characters
  - Estimate:
    - $p(x^t | x^{t-1}, \dots, x^1)$

prediction

memory

$\widehat{y^t} = g(Vs^t)$

$s^t = f(Ws^{t-1} + Ux^t)$



- Example
  - « S'il te plaît… dessine-moi …» what next ?
  - « $x^1 x^2 x^3$ … … … … … . … . . x^{t-1} … » what is x^t?

Background- - Language models – example

(Karpathy 2015- https://karpathy.github.io/2015/05/21/rnneffectiveness/)

- Training on Tolstoy's War and Peace a character language model
  - Stacked recurrent networks (LSTM)

```
tyntd-iafhatawiaoihrdemot  lytdws  e ,tfti, astai f ogoh eoase rrranbyne 'nhthnee e
plia tklrgd t o idoe ns,smtt   h ne etie h,hregtrs nigtike,aoaenns lng
```

train more

```
"Tmont thithey" fomesscerliund
Keushey. Thom here
sheulke, anmerenith ol sivh I lalterthend Bleipile shuwy fil on aseterlome
coaniogennc Phe lism thond hon at. MeiDimorotion in ther thize."
```

train more

```
Aftair fall unsuch that the hall for Prince Velzonski's that me of
her hearly, and behs to so arwage fiving were to it beloge, pavu say falling misfort
how, and Gogition is so overelical and ofter.
```

train more

```
"Why do what that day," replied Natasha, and wishing to himself the fact the
princess, Princess Mary was easier, fed in had oftened him.
Pierre aking his soul came to the packs and drove up his father-in-law women.
```

Background: Learning Grounded word representation

- Joint models
  - Learn jointly the representations from visual and textual inputs
    - LDA (a latent variable generates the two modalities)
    - Auto-encoders: 1 for each modality + a joint representation used e.g. for classification
    - Extensions of Word2Vec: e.g. perceptual information about a concrete concept is introduced in the text model whenever it is encounterd in text (e.g. Lazaridou et al. 2015)
- Sequential models
  - Learn separately text and image representations
  - Combine them
    - Middle fusion: e.g. concatenation of text and image representations, projections (e.g. SVD) or linear combinations of such concatenations
    - Late fusion: score combination of visual and textual representations, score are computed for a task (e.g. classification)
-

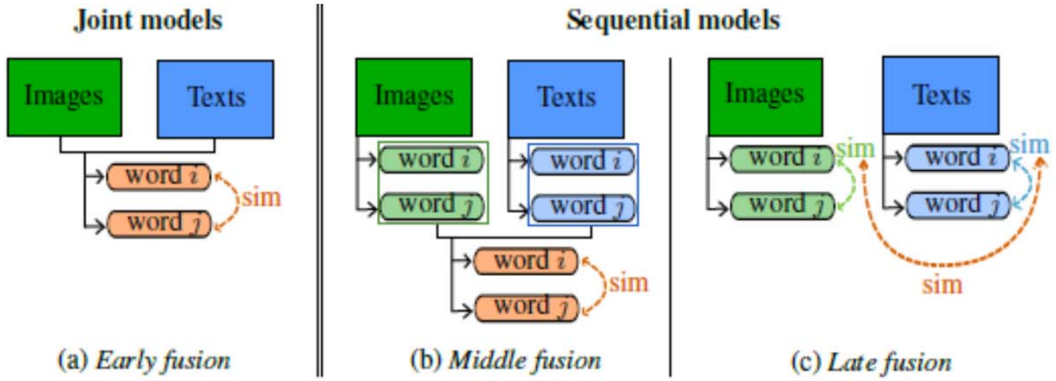


Figure 2: Overview of early fusion, middle fusion, and late fusion techniques. Round-corner rectangles denote word embeddings. Green is related to images and blue to text, orange round-corner rectangles are multimodal embeddings built from textual and visual resources. "sim" stands for an example of an evaluation task, namely *word similarity*.

Learning Multi-Modal Word Representation Grounded in Visual Context

E. Zablocki, B. Piwowarski, L. Soulier, P. Gallinari, AAAI 2018

Objective

- Context
  - Language is ambiguous, biased and lacks common sense
  - Images are unequivocal depictions
  - Human meaning representation is grounded in physical reality and sensorimotor experience
- Traditional Distributional Semantic models lack common sense
  - Word meaning link to physical world
  - Common sense
    - e.g. How do we know
      - the different functions of objects – bowls can hold soup, knifes are used to cut meat
      - the relations between objects (spatial, temporal), …
- Question
  - Can language be grounded in the visual world ?
    - Long term objective
  - Intermediate objective
    - Learn grounded representations for words

Importance of context

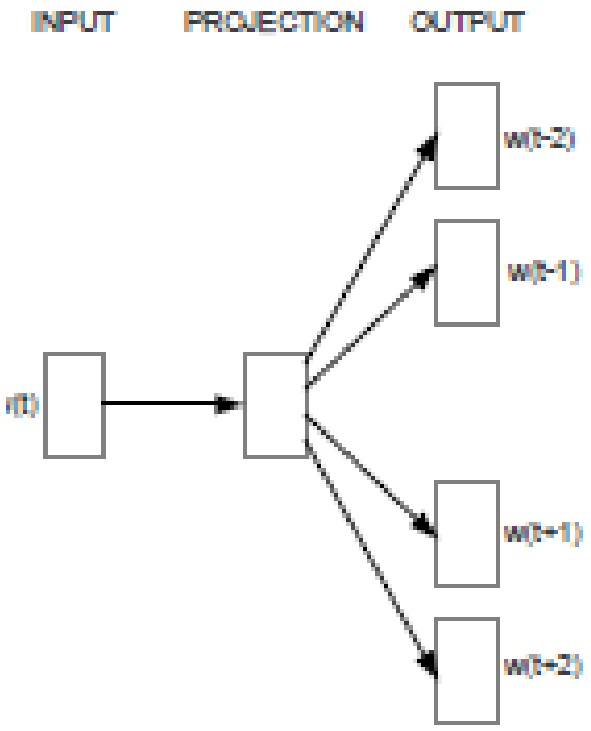
- Multimodal models show the complementarity of text and language
- They rely on direct image features + text alignment
- They ignore visual context
- Context brings a lot of information
  - How it is used, where it can be found, etc
- Approach
  - Hypothesis: visual context enhances the learned representation of words
  - Ground words in visual **context**
  - Without the need for text-image alignment
    - Trained on independent text / image corpora
    - Requires only that the 2 corpora share common entities
    - But contexts are different

Figure 3: Image of complex scene with lots of visual context for all objects

Background: Learning word vector representations Language model/ Skip Gram model (Mikolov et al. 2013)

- Predict the context word probability for a word in a sequence: $p(w_{\text{context}}|w_{input})$

- $p(w_{\text{context}}|w_{input}) = \frac{\exp(\boldsymbol{v}_{w_{\text{context}}}.\boldsymbol{v}_{w_{input}})}{\sum_{w=1}^{V}\exp(\boldsymbol{v}_{w}.\boldsymbol{v}_{w_{input}})}$
  - $\boldsymbol{v}_w$ is the learned representation of the w vector (the hidden layer), $\boldsymbol{v}_{w_{context}}.\boldsymbol{v}_{w_{\text{input}}}$ is a dot product and V is the vocabulary size



Skip-gram

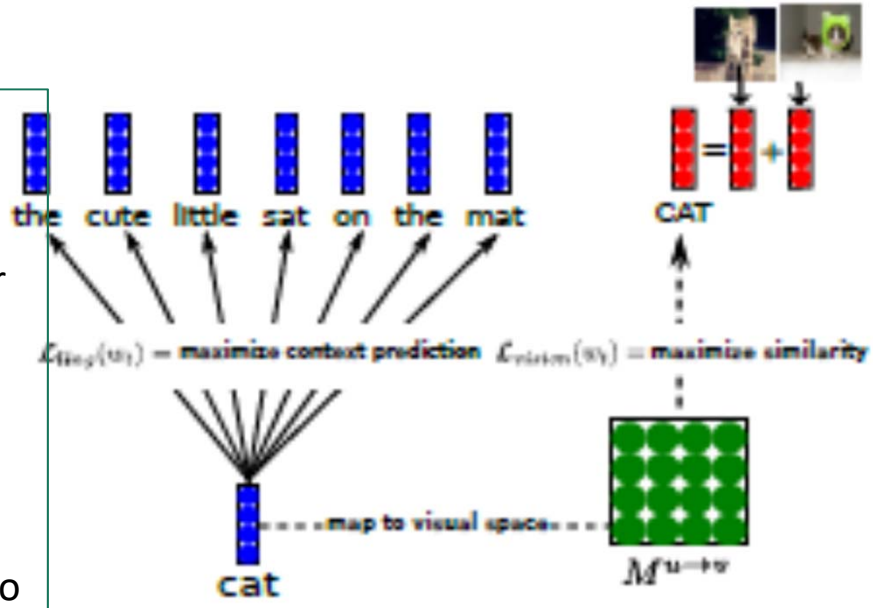
Background: Learning word vector representations Language model/ Skip Gram model (Mikolov et al. 2013)

- Loss function: average log probability
  - $L = \frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t)$
  - T is the number of words in the whole sequence used for training (roughly number of words in the corpus) and c is the context size
- Oin practice: simplified version of negative sampling
  - $l(w_{\text{input}}, w_{context}) = \log\sigma\left(v_{w_{context}}.v_{w_{input}}\right) + \sum_{i=1}^{k}\log\sigma(-v_{w_i}.v_{w_{input}}))$
  - With w_i a negative example sampled from the word distribution and $\sigma(x) = \frac{1}{1+\exp(-x)}$

Related work: Multi-modal Skip-Gram - Lazaridou et al. 2015

- Extension of skip-gram for multi-modal data
  - For a subset of the corpus words, textual corpus concepts are linked to visual representation of concepts
    - Representation of words corresponding to image entities are forced to be close to the visual pretrained representation of the corresponding object
    - Joint learning, No context

Figure from Lazaridou et al. 2015
- Left: text Skip-Gram
- Right: visual component
  - visual vectors obtained from the upper layer of a CNN for 5000 words occuring > 500 times in Imagenet (about 5% tokens in the text corpus)
  - Visual word: average vector learned from 100 samples of this item (e.g. cat)
  - M matrix: maps the textual representation to the visual one for this subset of words



Grounding words in visual context: research questions

- Research questions
  - RQ1: what is a visual context, what is a context model?
  - RQ2: how to learn joint representations from texts and images using the context
  - RQ3: how can we evaluate the contribution of the visual modality ti the final embeddings?

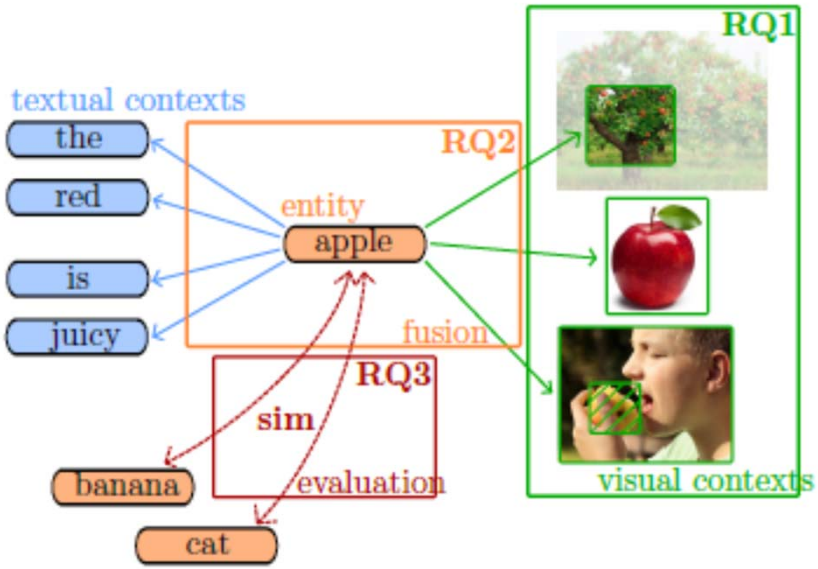


Figure 1: Illustrative figure of the 3 research questions raised in our paper. **RQ1**: How to define a visual context? **RQ2**: How to integrate visual context use in a multimodal model? **RQ3**: How to evaluate the visual contribution to learned embeddings

Context in visual data (RQ1)

- Entity e: object in the image
- Context element c: different possibilities corresponding to different « supervision » levels
  - High level context: surrounding objects
    - representation: learned embedding
  - Low level context: image patches
    - Full image with entity masked or random patch chosen around the entity
    - Representation: CNN encoding
  - Enhancing via spatial information
    - If entity localization in images is available, this information could be incorporated in the context representation
      - e.g. relative position of the bounding boxes of the object and the context

Model/ Loss functions (RQ2)

- Image context loss function
  - $L_{Image} = \sum_{e \in D} \sum_{c \in C_e} \left(\log \sigma(f_\theta(c).t_e) + \sum_{c^-} \log \sigma(-f_\theta(c^-).t_e) \right)$
    - $f_\theta(c) \in R^d$: representation of context c, D set of entities, t_e shared entity embedding
    - Similar to Skip-gram loss, except for functional $f_\theta(c)$

- Multimodal loss function
  - $L(T, U, \theta) = L_{text}(T, U) + \alpha L_{Image}(T, \theta)$
    - T: **shared** multi-modal text embeddings
    - U: **textual** context parameters
    - θ: **visual** context parameters

Experiments/ Tasks (RQ3)

- No direct evaluation - like for all entity learning methods

- Word pairs similarity
  - Datasets: human judgements on word pairs similarity
  - Compare human similarity scores with cosine similarity of learned entity representation
  - Measure: Spearman coefficient between the two lists of similarities

- Feature norm prediction
  - Feature norms are lists of attributes for an object
  - Predict object attributes from their representation (e.g. is_red, can_fly) (McRae et al. 2005: properties concrete nouns)
  - 43 characteristics grouped in 9 categories, 417 entities
  - Linear SVM classifier for the 9 categories on entity representation

- Concreteness prediction
  - Predict word concreteness score (gold ratings provided on 3260 words)
  - SVM regression on entity representation

Experiments/ Datasets

- Text
  - Dump of Wikipedia
    - 4.2 M articles, 2.1 unique words
- Visual data
  - Visual Genome (Krishna et al 2017)
  - 108 k images, 4842 unique entities, complex scenes, 31 object instances per image on average

Fig. 1: An overview of the data needed to move from perceptual awareness to cognitive understanding of images. We present a dataset of images densely annotated with numerous region descriptions, objects, attributes, and relationships. Some examples of region descriptions (e.g. "girl feeding large elephant" and "a man taking a picture behind girl") are shown (top). The objects (e.g. `elephant`), attributes (e.g. `large`) and relationships (e.g. `feeding`) are shown (bottom). Our dataset also contains image related question answer pairs (not shown).

Fig: Krishna et al 2017

Experiments/ Models

- Baselines
  - Text alone
  - Text + direct visual features from objects
  - Text + visual context (sequential models)
- Model
  - Objects
  - Patches (Full image, random patches)
  - Spatial information
  - Ensemble model: context + direct visual features

Experiments/ results

- Example results for RQ2

Word similarity
Spearman correlation
4 datasets

Feature norm prediction
f1-scores
9 categories

Concreteness

			VisSim	SemSim	Simlex	MEN	WordSim	Encyclopedic	Taste	Sound	Taxonomic	Function	Tactile	Color	Shape	Motion	Conc.
			Similarity Evaluation					Feature-norm Prediction Task									Conc.
Basel.	Text	$\mathbf{T}$	48	60	33	69	63	58	52	44	79	62	11	32	54	60	42.1
	Sequential	$\mathbf{O} \oplus \mathbf{T}$	49	62	33	71	64	63	55	40	72	59	12	35	54	58	43.7
	Joint	$\mathbf{L} + \mathbf{T}$	52	65	34	71	65	61	55	42	80	59	11	31	54	62	43.4
Our models	Objects	$\mathbf{O} + \mathbf{T}$	53	66	35	**75**	**67**	62	55	46	**82**	61	13	33	55	61	42.9
	Patches	$\mathbf{P} + \mathbf{T}$	53	65	35	72	**67**	60	56	49	**82**	60	12	32	55	61	43.1
		$\mathbf{P}_{\text{full}} + \mathbf{T}$	53	65	34	73	65	60	55	44	**82**	**63**	14	32	55	59	43.2
	Spatial	$\mathbf{Sp}(\mathbf{O}, \delta, \oplus) + \mathbf{T}$	52	66	36	73	64	**64**	**59**	46	81	62	06	31	**57**	**63**	42.5
		$\mathbf{Sp}(\mathbf{O}, c, \oplus) + \mathbf{T}$	54	66	35	72	64	62	56	**52**	80	61	13	**34**	**57**	58	43.7
		$\mathbf{Sp}(\mathbf{O}, \delta, b) + \mathbf{T}$	54	**68**	**38**	73	66	63	56	48	81	60	13	32	56	**63**	42.5
		$\mathbf{Sp}(\mathbf{O}, c, b) + \mathbf{T}$	**55**	67	34	**75**	64	61	58	46	80	**63**	**15**	**34**	**57**	62	**44.4**
	Ensemble	$\mathbf{L} + \mathbf{O} + \mathbf{T}$	54	66	35	**75**	65	63	55	50	**82**	60	10	33	55	59	43.9

Table 2: RQ2 experimental results on word similarity evaluation benchmarks, feature-norm prediction task, concreteness prediction task (Conc.). Concreteness measures are coefficients of determination (R^2) multiplied by 100.

Experiments/ results

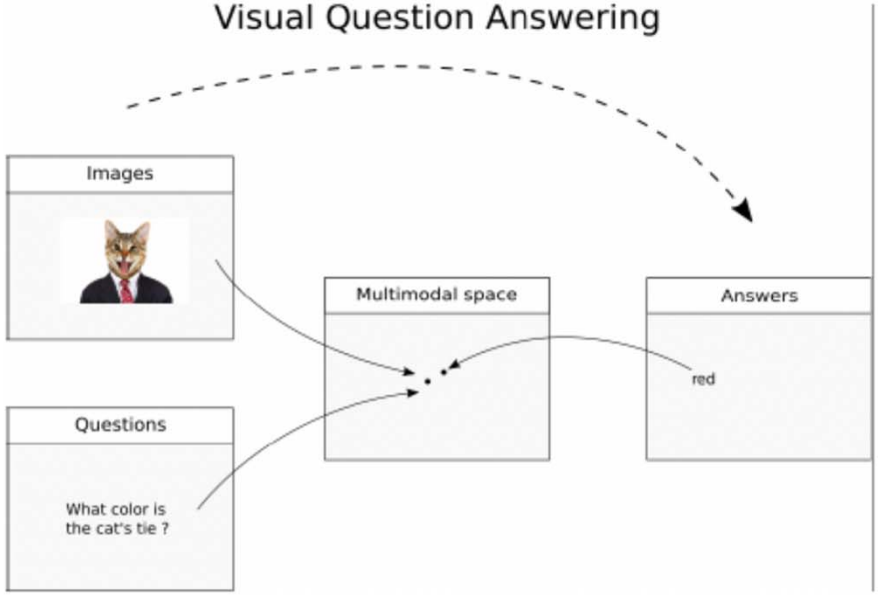
- Main results on all three tasks/ RQ
  - Multi-modal embeddings > Text only embeddings
  - Surroundings > Direct features (word sim. task)
  - Ensemble model: surrounding and direct features complementary
  - Spatial information useful
  - High-level context > Low-level context

Multimodal Tucker Fusion for Visual Question Answering

H. Ben Younes, R. Cadene, M. Cord, N. Thome, ICCV 2017

Visual Question Answering

- Classification using multimodal inputs
- several datasets



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

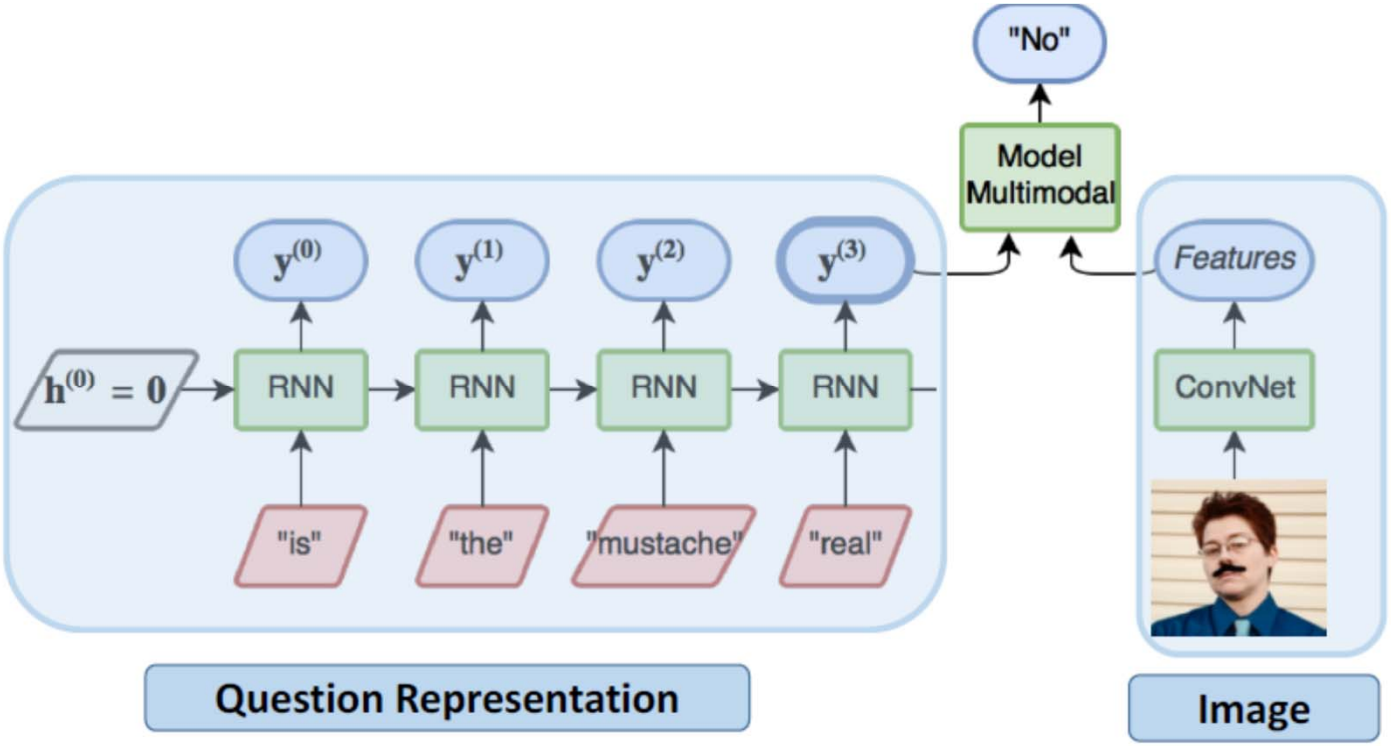
How many slices of pizza are there?
Is this a vegetarian pizza?

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)

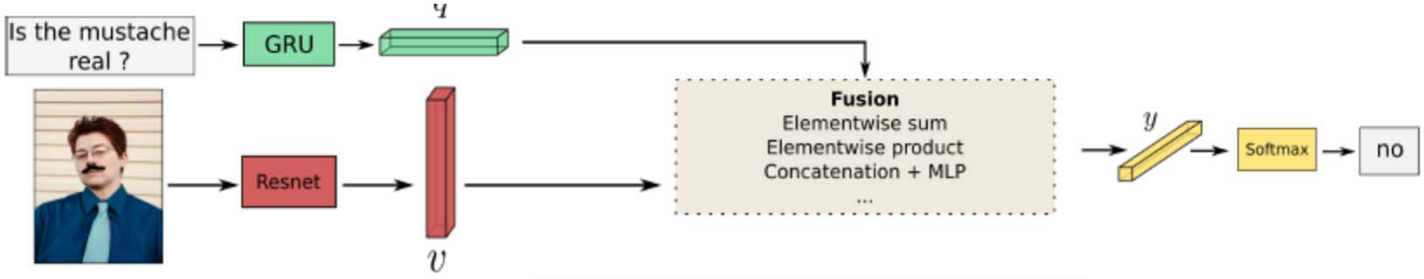
Vanilla VQA Scheme

•



VQA: Multimodal fusion strategies

- Linear models



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

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

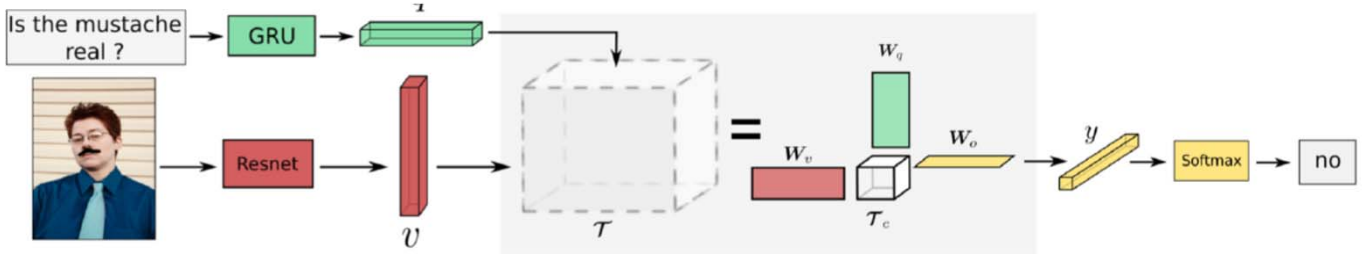
$$\text{Element-wise product} : y = (\boldsymbol{W}\mathbf{q}) \odot (\boldsymbol{V}\mathbf{v})$$

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

Works best among linear methods

VQA: multimodal fusion strategies

- Bilinear models: extract more relevant correlations between the two modes



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

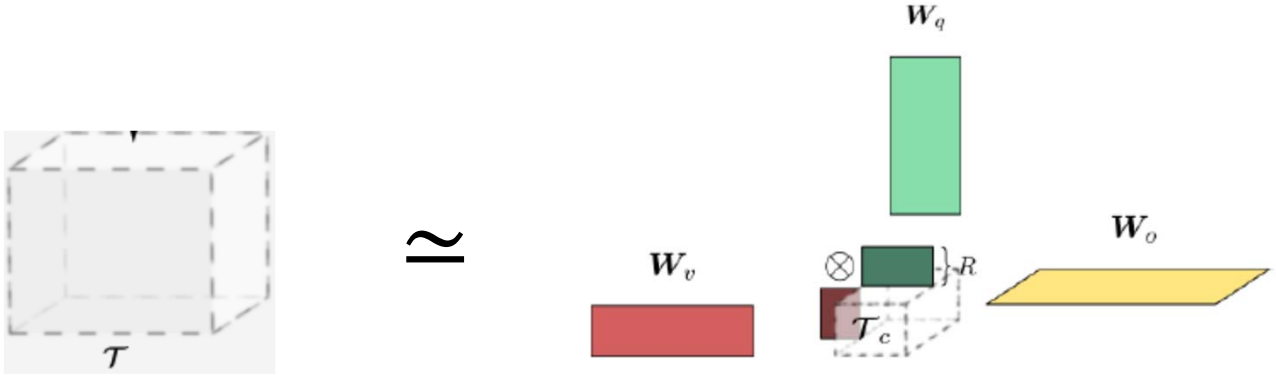
$$\mathbf{y}^k = \sum_{i=1}^{d_q} \sum_{j=1}^{d_v} \mathcal{T}^{ijk} \mathbf{q}^i \mathbf{v}^j$$

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

- Full tensor: $d_q = d_v = 2048$ and 2000 classes, number of free parameters in $T \sim 10^{10}$
- Reduce the size of the tensor

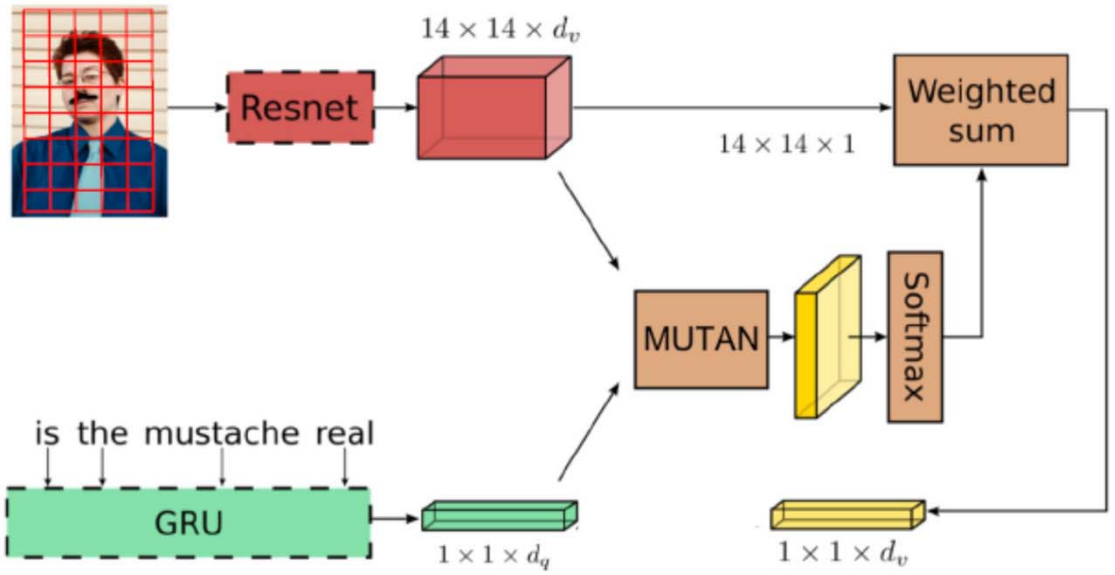
VQA: multimodal fusion strategies

- Reducing tensor size
  - Tucker based decomposition + sparsity constraints on the core tensor T_c

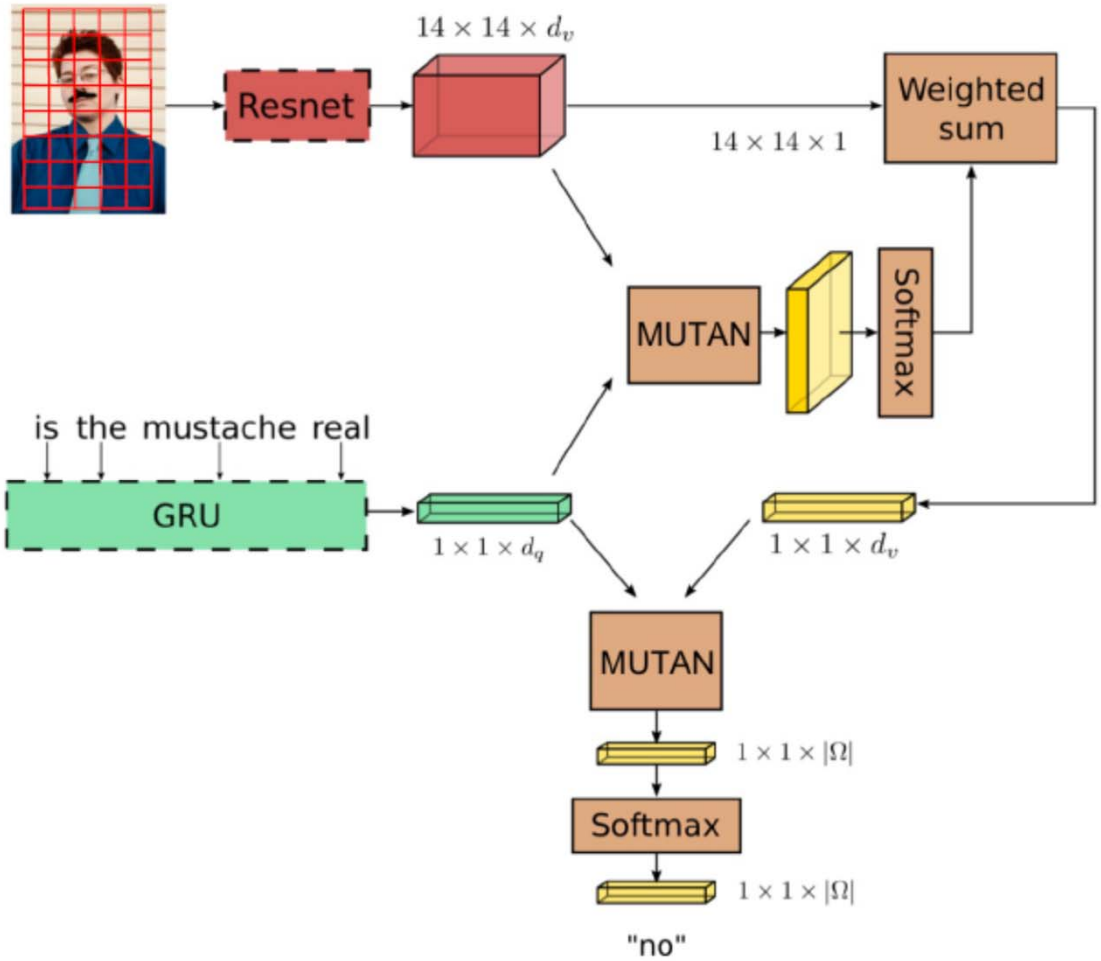


  - All the parameters (W_v, W_q, W_o, T_c) are learned end to end

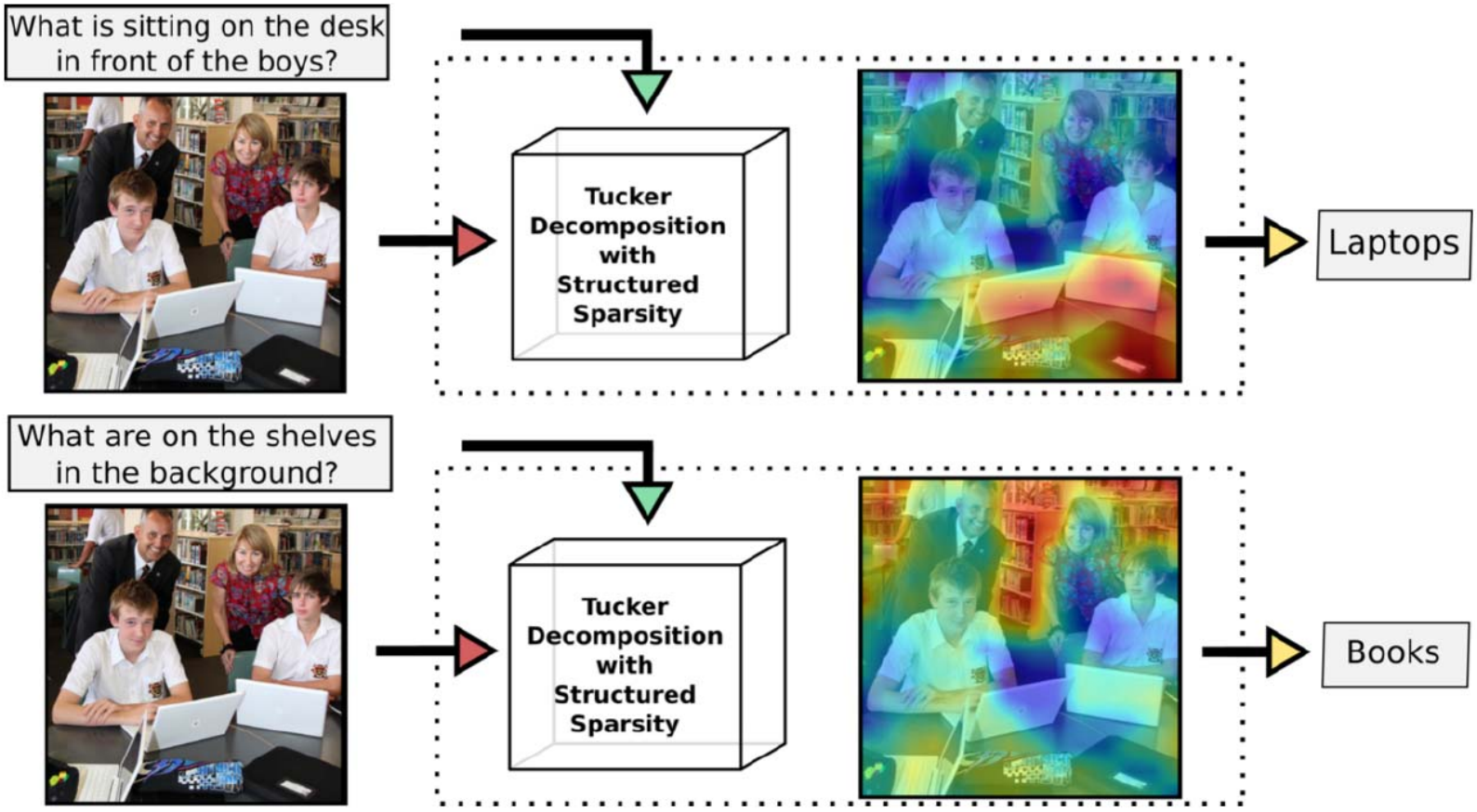
VQA: attention process

- 

VQA: attention process

- 

Attention mechanism



VQA and 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]

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**.