Merging language and vision modalities: Last years work

Raffaella Bernardi

University of Trento

November, 2017

Last time

Last time we have introduced the first computational work on Language and Vision integration.
Today, we look at new tasks that have been proposed more recently.

Layout

1. Cross Modal Mapping
2. Visual Phrases
3. Tasks
4. Intermezzo
5. Find their limitations
6. New task: Visual Reasoning
7. Others
8. Conclusion

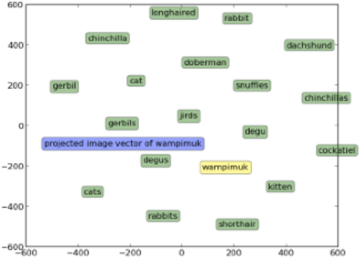
Cross-modal mapping: Generalization

Angeliki Lazaridou, Elia Bruni and Marco Baroni. (ACL 2014)
Transfering knowledge acquired in one modality to the other one.
Learn to project one space into the other, from the visual space onto the language space. Two tasks:

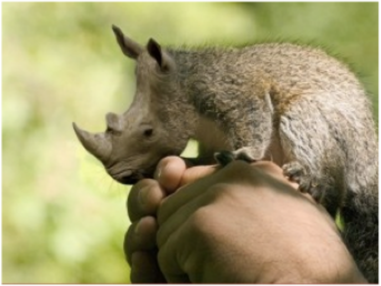
- **Zero-Shot Learning**
- **Fast Mapping**

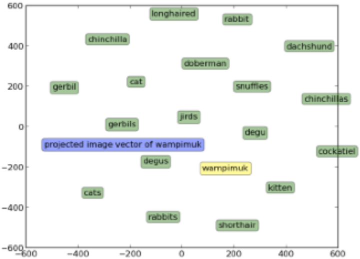
In both tasks, the projected vector of the unseen concept is labeled with the word associated to its cosine-based nearest neighbor vector in the corresponding semantic space.

Zero-Shot Learning: the task



Zero-Shot Learning: the task





Zero-Shot Learning

Learn a classifier $X \rightarrow Y$, s.t. X are images, Y are language vectors. Label an image of an unseen concept with the word associated to its cosine-based nearest neighbor vector in the language space.

For a subset of concepts (e.g., a set of animals, a set of vehicles), we possess information related to both their linguistic and visual representations.

- During training, this cross-modal vocabulary is used to induce a projection function, which intuitively represents a mapping between visual and linguistic dimensions.
  Thus, this function, given a visual vector, returns its corresponding linguistic representation.
- At test time, the system is presented with a previously unseen object (e.g., wampimuk). This object is projected onto the linguistic space and associated with the word label of the nearest neighbor in that space (containing all the unseen and seen concepts).

Zero-Shot Learning

Learn a classifier $X \rightarrow Y$, s.t. X are images, Y are language vectors. Label an image of an unseen concept with the word associated to its cosine-based nearest neighbor vector in the language space.
For a subset of concepts (e.g., a set of animals, a set of vehicles), we possess information related to both their linguistic and visual representations.

- During training, this cross-modal vocabulary is used to induce a projection function, which intuitively represents a mapping between visual and linguistic dimensions.
  Thus, this function, given a visual vector, returns its corresponding linguistic representation.
- At test time, the system is presented with a previously unseen object (e.g., wampimuk). This object is projected onto the linguistic space and associated with the word label of the nearest neighbor in that space (containing all the unseen and seen concepts).

Zero-Shot Learning

Learn a classifier $X \rightarrow Y$, s.t. X are images, Y are language vectors. Label an image of an unseen concept with the word associated to its cosine-based nearest neighbor vector in the language space.
For a subset of concepts (e.g., a set of animals, a set of vehicles), we possess information related to both their linguistic and visual representations.

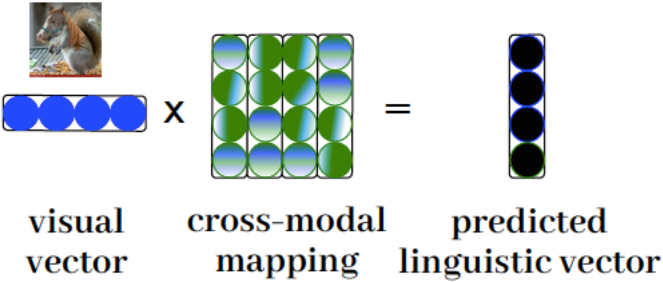
- During training, this cross-modal vocabulary is used to induce a projection function, which intuitively represents a mapping between visual and linguistic dimensions.
  Thus, this function, given a visual vector, returns its corresponding linguistic representation.
- At test time, the system is presented with a previously unseen object (e.g., wampimuk). This object is projected onto the linguistic space and associated with the word label of the nearest neighbor in that space (containing all the unseen and seen concepts).

Zero-Shot Learning

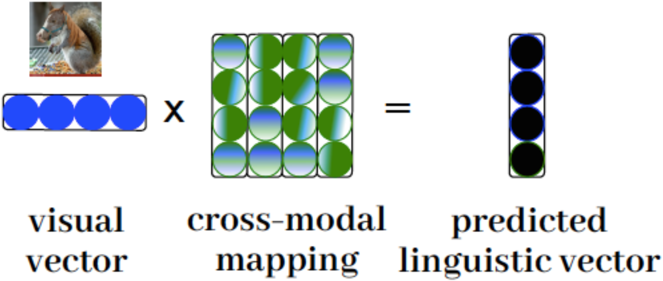
Learn a classifier $X \rightarrow Y$, s.t. X are images, Y are language vectors. Label an image of an unseen concept with the word associated to its cosine-based nearest neighbor vector in the language space.
For a subset of concepts (e.g., a set of animals, a set of vehicles), we possess information related to both their linguistic and visual representations.

- During training, this cross-modal vocabulary is used to induce a projection function, which intuitively represents a mapping between visual and linguistic dimensions.
  Thus, this function, given a visual vector, returns its corresponding linguistic representation.
- At test time, the system is presented with a previously unseen object (e.g., wampimuk). This object is projected onto the linguistic space and associated with the word label of the nearest neighbor in that space (containing all the unseen and seen concepts).

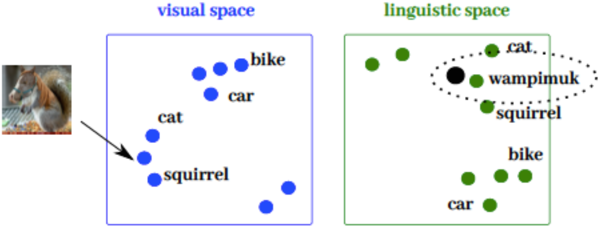
Zero-shot leaning: linear mapping



Zero-shot leaning: linear mapping



Zero-shot leaning: example



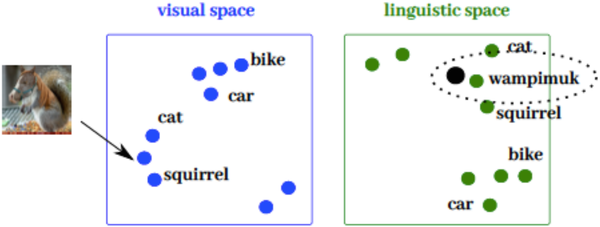
Step 1 Obtain "**parallel data**" of linguistic and visual vectors of concepts.

Step 2 Learn a cross-modal mapping between the two semantic spaces

Step 3 Map the **unknown** concept onto the linguistic/visual space

Step 4 Obtain a label through **nearest neighbor search**

Zero-shot leaning: example



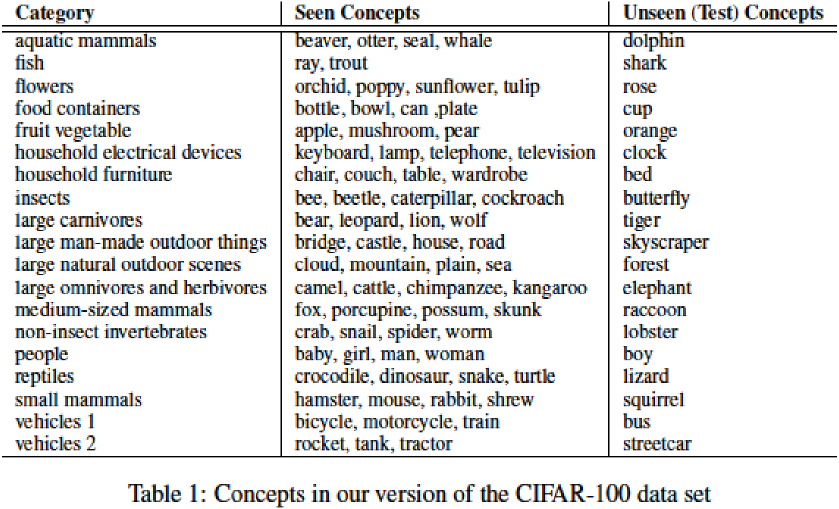
Step 1 Obtain "**parallel data**" of linguistic and visual vectors of concepts.

Step 2 Learn a cross-modal mapping between the two semantic spaces

Step 3 Map the **unknown** concept onto the linguistic/visual space

Step 4 Obtain a label through **nearest neighbor search**

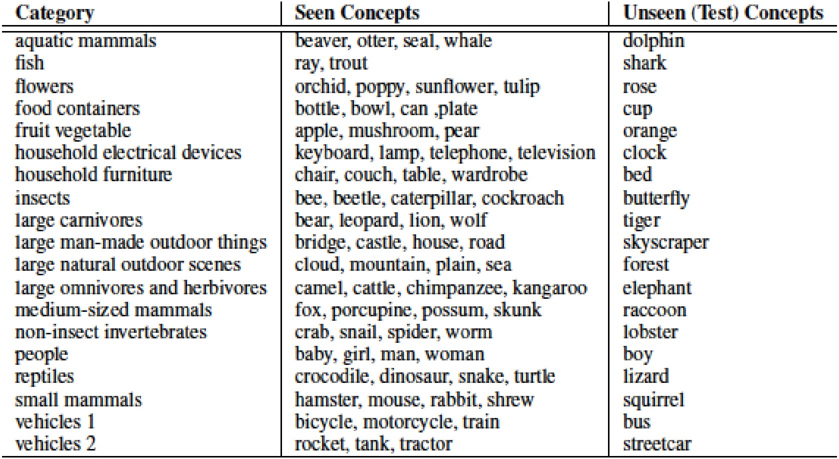
Dataset



Category	Seen Concepts	Unseen (Test) Concepts
aquatic mammals	beaver, otter, seal, whale	dolphin
fish	ray, trout	shark
flowers	orchid, poppy, sunflower, tulip	rose
food containers	bottle, bowl, can ,plate	cup
fruit vegetable	apple, mushroom, pear	orange
household electrical devices	keyboard, lamp, telephone, television	clock
household furniture	chair, couch, table, wardrobe	bed
insects	bee, beetle, caterpillar, cockroach	butterfly
large carnivores	bear, leopard, lion, wolf	tiger
large man-made outdoor things	bridge, castle, house, road	skyscraper
large natural outdoor scenes	cloud, mountain, plain, sea	forest
large omnivores and herbivores	camel, cattle, chimpanzee, kangaroo	elephant
medium-sized mammals	fox, porcupine, possum, skunk	raccoon
non-insect invertebrates	crab, snail, spider, worm	lobster
people	baby, girl, man, woman	boy
reptiles	crocodile, dinosaur, snake, turtle	lizard
small mammals	hamster, mouse, rabbit, shrew	squirrel
vehicles 1	bicycle, motorcycle, train	bus
vehicles 2	rocket, tank, tractor	streetcar

Table 1: Concepts in our version of the CIFAR-100 data set

Dataset

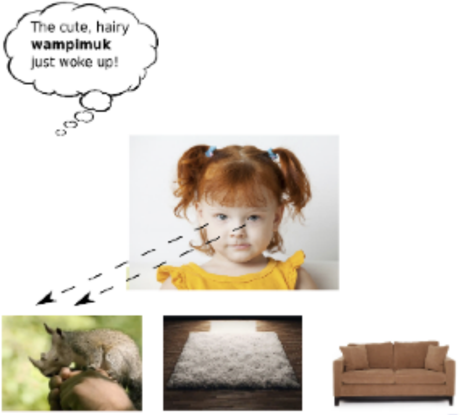


Category	Seen Concepts	Unseen (Test) Concepts
aquatic mammals	beaver, otter, seal, whale	dolphin
fish	ray, trout	shark
flowers	orchid, poppy, sunflower, tulip	rose
food containers	bottle, bowl, can ,plate	cup
fruit vegetable	apple, mushroom, pear	orange
household electrical devices	keyboard, lamp, telephone, television	clock
household furniture	chair, couch, table, wardrobe	bed
insects	bee, beetle, caterpillar, cockroach	butterfly
large carnivores	bear, leopard, lion, wolf	tiger
large man-made outdoor things	bridge, castle, house, road	skyscraper
large natural outdoor scenes	cloud, mountain, plain, sea	forest
large omnivores and herbivores	camel, cattle, chimpanzee, kangaroo	elephant
medium-sized mammals	fox, porcupine, possum, skunk	raccoon
non-insect invertebrates	crab, snail, spider, worm	lobster
people	baby, girl, man, woman	boy
reptiles	crocodile, dinosaur, snake, turtle	lizard
small mammals	hamster, mouse, rabbit, shrew	squirrel
vehicles 1	bicycle, motorcycle, train	bus
vehicles 2	rocket, tank, tractor	streetcar

Table 1: Concepts in our version of the CIFAR-100 data set

Cross Modal Mapping

Fast Mapping



Fast Mapping

Learn a word vector from a *few sentences*, associate it to the referring image exploiting cosine-based neighbor vector in the visual space.

The fast mapping setting can be seen as a special case of the zero-shot task. Whereas for the latter our system assumes that all concepts have rich linguistic representations (i.e., representations estimated from a large corpus), in the case of the former, new concepts are assumed to be encounted in a limited linguistic context and therefore lacking rich linguistic representations.

This is operationalized by constructing the text-based vector for these concepts from a context of just a few occurrences. In this way, we simulate the first encounter of a learner with a concept that is new in both visual and linguistic terms.

New paper: *Multimodal semantic learning from child-directed input* Angeliki Lazaridou, Grzegorz Chrupala, Raquel Fernandez and Marco Baroni NAACL 2016 Short `http://clic.cimec.unitn.it/marco/publications/lazaridou-etal-multimodal-learning-from-cdi-naacl2016.pdf`

Fast Mapping

Learn a word vector from a *few sentences*, associate it to the referring image exploiting cosine-based neighbor vector in the visual space.
The fast mapping setting can be seen as a special case of the zero-shot task. Whereas for the latter our system assumes that all concepts have rich linguistic representations (i.e., representations estimated from a large corpus), in the case of the former, new concepts are assumed to be encounted in a limited linguistic context and therefore lacking rich linguistic representations.
This is operationalized by constructing the text-based vector for these concepts from a context of just a few occurrences. In this way, we simulate the first encounter of a learner with a concept that is new in both visual and linguistic terms.

New paper: *Multimodal semantic learning from child-directed input* Angeliki Lazaridou, Grzegorz Chrupala, Raquel Fernandez and Marco Baroni NAACL 2016 Short `http://clic.cimec.unitn.it/marco/publications/lazaridou-etal-multimodal-learning-from-cdi-naacl2016.pdf`

Fast Mapping

Learn a word vector from a *few sentences*, associate it to the referring image exploiting cosine-based neighbor vector in the visual space.
The fast mapping setting can be seen as a special case of the zero-shot task. Whereas for the latter our system assumes that all concepts have rich linguistic representations (i.e., representations estimated from a large corpus), in the case of the former, new concepts are assumed to be encounted in a limited linguistic context and therefore lacking rich linguistic representations.
This is operationalized by constructing the text-based vector for these concepts from a context of just a few occurrences. In this way, we simulate the first encounter of a learner with a concept that is new in both visual and linguistic terms.

New paper: *Multimodal semantic learning from child-directed input* Angeliki Lazaridou, Grzegorz Chrupala, Raquel Fernandez and Marco Baroni NAACL 2016
Short `http://clic.cimec.unitn.it/marco/publications/lazaridou-etal-multimodal-learning-from-cdi-naacl2016.pdf`

Layout

1. Cross Modal Mapping
2. Visual Phrases
3. Tasks
4. Intermezzo
5. Find their limitations
6. New task: Visual Reasoning
7. Others
8. Conclusion

Images as Visual Phrases

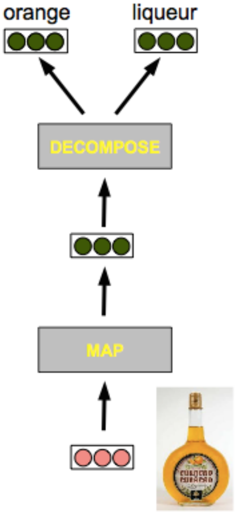
- Given the visual representation of an object, can we “decompose” it into attribute and object?
- Can we learn the visual representation of attributes and learn to compose them with the visual representation of an object?

Visual Phrase: Decomposition

A. Lazaridou, G. Dinu, A. Liska, M. Baroni (TACL 2015)

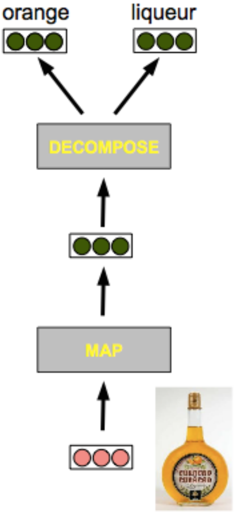
- First intuition: vision and language space have similar structures (also w.r.t attribute/adjectives)
- Second intuition: Objects are bundles of attributes. Hence, attributes are implicitely learned together with objects.

Decomposition Model: attribute annotation



Evaluation: (unseen) object/noun and attribute/adjective retrieval.

Decomposition Model: attribute annotation



Evaluation: (unseen) object/noun and attribute/adjective retrieval.

Images as Visual Phrases: Composition

Coloring Objects: Adjective-Noun Visual Semantic Compositionality (VL'14) D.T. Nguyen, A. Lazaridou and R. Bernardi

1. Assumption from linguistics: Adjectives are noun modifiers. They are functions from N into N.
2. From COMPOSES: adjectives can be learned from (ADJ N, N) inputs.
3. Applied to images: Compositional Visual Model?

Images as Visual Phrases: Composition

Coloring Objects: Adjective-Noun Visual Semantic Compositionality (VL'14) D.T. Nguyen, A. Lazaridou and R. Bernardi

1. Assumption from linguistics: Adjectives are noun modifiers. They are functions from N into N.
2. From COMPOSES: adjectives can be learned from (ADJ N, N) inputs.
3. Applied to images: Compositional Visual Model?

Images as Visual Phrases: Composition

Coloring Objects: Adjective-Noun Visual Semantic Compositionality (VL'14) D.T. Nguyen, A. Lazaridou and R. Bernardi

1. Assumption from linguistics: Adjectives are noun modifiers. They are functions from N into N.
2. From COMPOSES: adjectives can be learned from (ADJ N, N) inputs.
3. Applied to images: Compositional Visual Model?

Visual Composition

From the visual representation:

- Dense-Sift feature vectors as Noun vectors (e.g. car. light)
- Color-Sift feature vectors as Phrase vectors (e.g. red car. red light)

Learn the function (color) that maps the noun to the phrase. Apply that function to new (unseen) objects (e.g. red truck) and retrieve the image. We compare the composed visual vector (ATT OBJ) vs. composed linguistic vectors (ADJ N) vs. observed linguistic vectors.

New paper: Misra et al. *From red wine to red tomato: Composition with Context* CVPR 2017

Visual Composition

From the visual representation:

- Dense-Sift feature vectors as Noun vectors (e.g. car. light)
- Color-Sift feature vectors as Phrase vectors (e.g. red car. red light)

Learn the function (color) that maps the noun to the phrase. Apply that function to new (unseen) objects (e.g. red truck) and retrieve the image. We compare the composed visual vector (ATT OBJ) vs. composed linguistic vectors (ADJ N) vs. observed linguistic vectors.

New paper: Misra et al. *From red wine to red tomato: Composition with Context* CVPR 2017

Visual Composition

From the visual representation:

- Dense-Sift feature vectors as Noun vectors (e.g. car. light)
- Color-Sift feature vectors as Phrase vectors (e.g. red car. red light)

Learn the function (color) that maps the noun to the phrase. Apply that function to new (unseen) objects (e.g. red truck) and retrieve the image. We compare the composed visual vector (ATT OBJ) vs. composed linguistic vectors (ADJ N) vs. observed linguistic vectors.

New paper: Misra et al. *From red wine to red tomato: Composition with Context* CVPR 2017

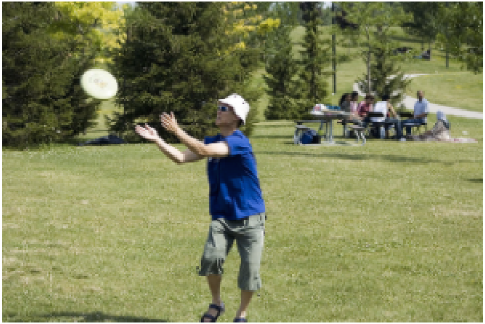
Layout

1. Cross Modal Mapping
2. Visual Phrases
3. Tasks
4. Intermezzo
5. Find their limitations
6. New task: Visual Reasoning
7. Others
8. Conclusion

New evaluation tasks

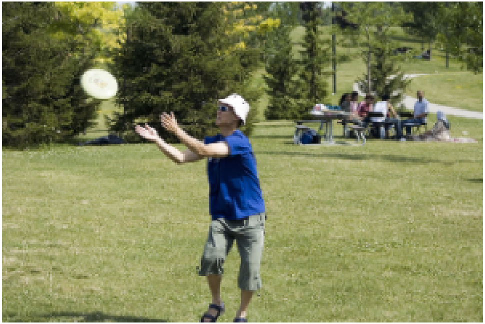
- Image Captioning (IC)
- Visual Question Answering (VQA)
- Visual Reasoning

Image Captioning (IC)



a man is throwing a frisbee in a park

Image Captioning (IC)



a man is throwing a frisbee in a park

IC: Overview

- **Datasets** Flickr, Pascal, MS-COCO (164K images, 5 captions each)
- **Survey** Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures, Bernardi et al. JAIR 2016
- **Very good talk** by Karpathy (2015): `https://www.youtube.com/watch?v=ZkY7fAoaNcg`

IC: Overview

- **Datasets** Flickr, Pascal, MS-COCO (164K images, 5 captions each)
- **Survey** Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures, Bernardi et al. JAIR 2016
- **Very good talk** by Karpathy (2015): `https://www.youtube.com/watch?v=ZkY7fAoaNcg`

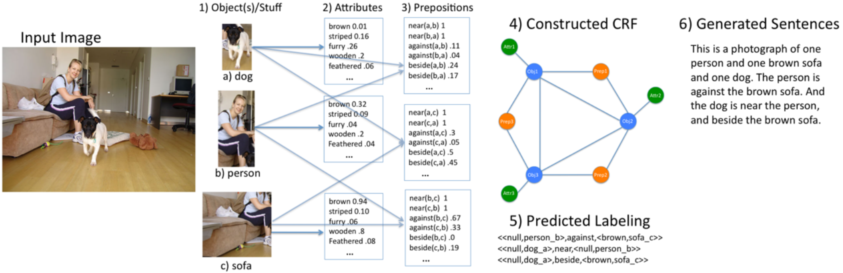
IC: Approaches

- Approaches: Retrieve vs. Generate
- Frameworks: Pipeline of predictions vs. End-to-end

IC approaches

Pipeline

E.g., Kulkarni et al. (2011)



Input Image

1) Object(s)/Stuff

a) dog

b) person

c) sofa

2) Attributes

brown 0.01
striped 0.16
furry .26
wooden .2
feathered .06
...

brown 0.32
striped 0.09
furry .04
wooden .2
Feathered .04
...

brown 0.94
striped 0.10
furry .06
wooden .8
Feathered .08
...

3) Prepositions

near(a,b) 1
near(b,a) 1
against(a,b) .11
against(b,a) .04
beside(a,b) .24
beside(b,a) .17
...

near(a,c) 1
near(c,a) 1
against(a,c) .3
against(c,a) .05
beside(a,c) .5
beside(c,a) .45
...

near(b,c) 1
near(c,b) 1
against(b,c) .67
against(c,b) .33
beside(b,c) .0
beside(c,b) .19
...

4) Constructed CRF

5) Predicted Labeling

<<null,person_b>,against,<brown,sofa_c>>
<<null,dog_a>,near,<null,person_b>>
<<null,dog_a>,beside,<brown,sofa_c>>

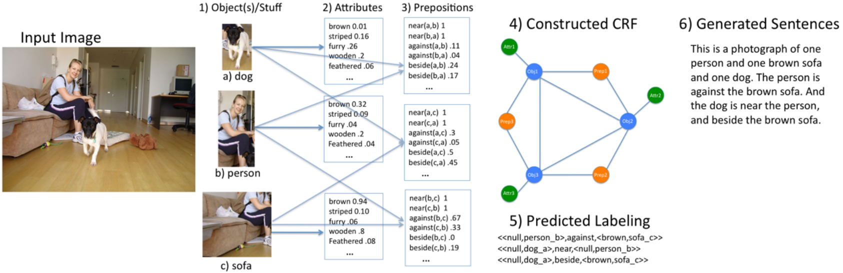
6) Generated Sentences

This is a photograph of one person and one brown sofa and one dog. The person is against the brown sofa. And the dog is near the person, and beside the brown sofa.

IC approaches

Pipeline

E.g., Kulkarni et al. (2011)



<<null,person_b>,against,<brown,sofa_c>>
<<null,dog_a>,near,<null,person_b>>
<<null,dog_a>,beside,<brown,sofa_c>>

This is a photograph of one person and one brown sofa and one dog. The person is against the brown sofa. And the dog is near the person, and beside the brown sofa.

IC approaches

End-to-end

E.g., Karpathy and Fei Fei (2015)

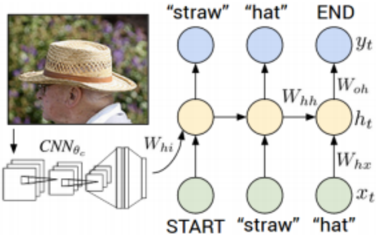


Figure 4. Diagram of our multimodal Recurrent Neural Network generative model. The RNN takes a word, the context from previous time steps and defines a distribution over the next word in the sentence. The RNN is conditioned on the image information at the first time step. START and END are special tokens.

IC approaches

End-to-end

E.g., Karpathy and Fei Fei (2015)

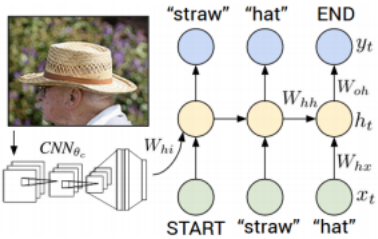


Figure 4. Diagram of our multimodal Recurrent Neural Network generative model. The RNN takes a word, the context from previous time steps and defines a distribution over the next word in the sentence. The RNN is conditioned on the image information at the first time step. START and END are special tokens.

IC: limitations

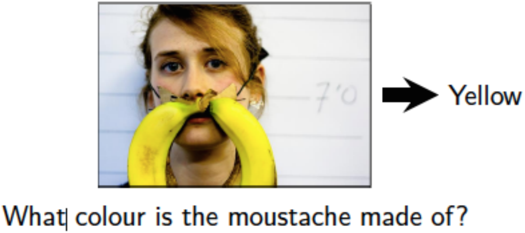
- Evaluation Measures: Bleu, Rouge, etc. but not precise.
- No reasoning

IC: limitations

- Evaluation Measures: Bleu, Rouge, etc. but not precise.
- No reasoning

Visual Question Answering (VQA)

VQA: Visual Question Answering Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, Devi Parikh (2016)

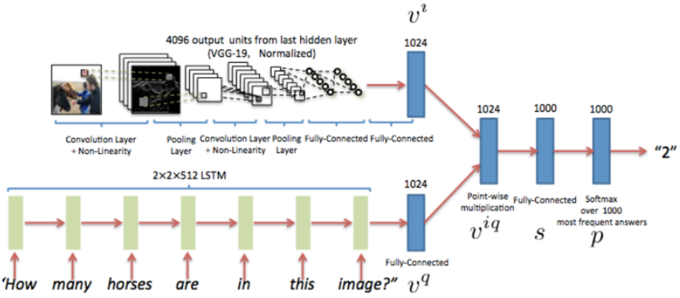


What colour is the moustache made of?

VQA: Overview

- **Datasets** DAQUAR 2014, COCO-QA, VQA, Visual7W, Visual Genome.
- **Survey** Visual Question Answering: A Survey of Methods and Datasets Wu et ali, (2016)

VQA: Model


v^i
4096 output units from last hidden layer
(VGG-19, Normalized)
1024
Convolution Layer + Non-Linearity
Pooling Layer
Convolution Layer + Non-Linearity
Pooling Layer
Fully-Connected
Fully-Connected
1024
1000
1000
"2"
2×2×512 LSTM
1024
Point-wise multiplication
v^{iq}
Fully-Connected
s
Softmax over 1000 most frequent answers
p
Fully-Connected
"How many horses are in this image?"
v^q

$$v^{iq} = v^i \circ v^q \qquad s = W v^{iq} + b \qquad p_a = \frac{e^{s_a}}{\sum_{a'} e^{s_{a'}}}$$

VQA

Limitations

- Language prior problem: Blind models perform pretty well (50% accuracy on COCO-VQA!).
- Development of new real image datasets: VQA2, FOIL, TDIUC, NLRV
- Development of synthetic datasets: SHAPES, CLEVR, Yin and Yang.

VQA

Limitations

- Language prior problem: Blind models perform pretty well (50% accuracy on COCO-VQA!).
- Development of new real image datasets: VQA2, FOIL, TDIUC, NLRV
- Development of synthetic datasets: SHAPES, CLEVR, Yin and Yang.

VQA

Limitations

- Language prior problem: Blind models perform pretty well (50% accuracy on COCO-VQA!).
- Development of new real image datasets: VQA2, FOIL, TDIUC, NLRV
- Development of synthetic datasets: SHAPES, CLEVR, Yin and Yang.

Best performing VQA model now

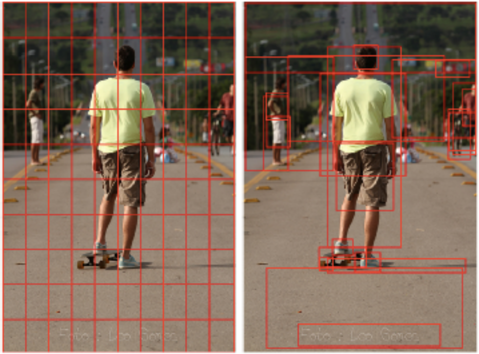


Figure 1. Typically, attention models operate on CNN features corresponding to a uniform grid of equally-sized image regions (left). Our approach enables attention to be calculated at the level of objects and other salient image regions (right).

Layout

1. Cross Modal Mapping
2. Visual Phrases
3. Tasks
4. Intermezzo
5. Find their limitations
6. New task: Visual Reasoning
7. Others
8. Conclusion

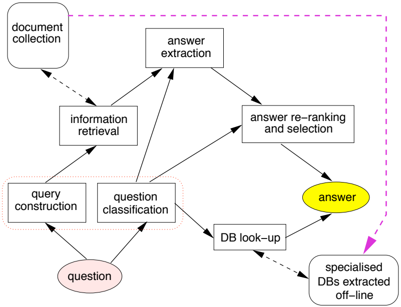
Question Answering

History

- In 1965 Simmons reviewed 15 QA systems.
- In the 60'/70' work on QA as front-end to databases. In the '90 decrease of interest in NLIDB. Nowadays: NLIDB on controlled natural language.
- In the '90 the rise of search engines brought QA over unstructured data to re-emerge. TREC QA track.
- Challanges of incremental difficulties (the answer is in the document, the answer is not in the document, the answer is spread in various documents; factual question, other types of questions etc.) Pipe-lines of modules.

Question Answering

Sample of QA pipe-line architecture



In the last year: boom of end-to-end systems; but let's not forget the ideas proposed in the past.

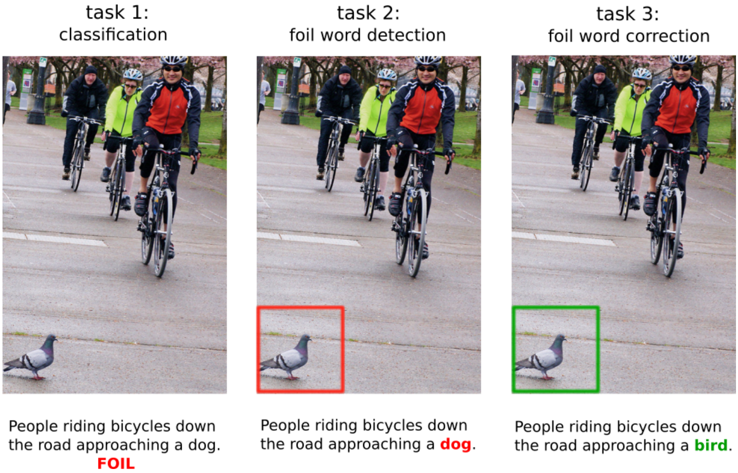
Layout

1. Cross Modal Mapping
2. Visual Phrases
3. Tasks
4. Intermezzo
5. Find their limitations
6. New task: Visual Reasoning
7. Others
8. Conclusion

Happy?

After the great success on IC and VQA people have started proposing tasks to highlight the *weakness* of current Language and Vision models.

FOIL



task 1:
classification

People riding bicycles down the road approaching a dog.
FOIL

task 2:
foil word detection

People riding bicycles down the road approaching a **dog**.

task 3:
foil word correction

People riding bicycles down the road approaching a **bird**.

One image associated with very similar captions but one is T and the other is F.

Irrelevant questions

Mahendru et al. EMNLP 2017

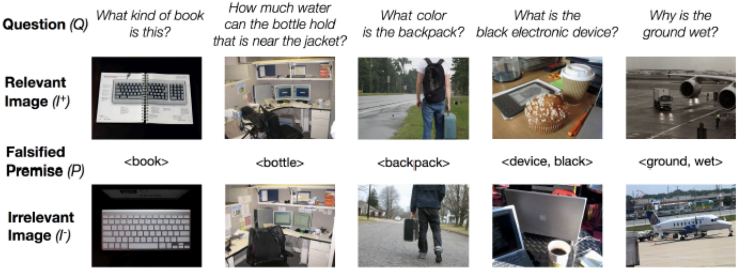
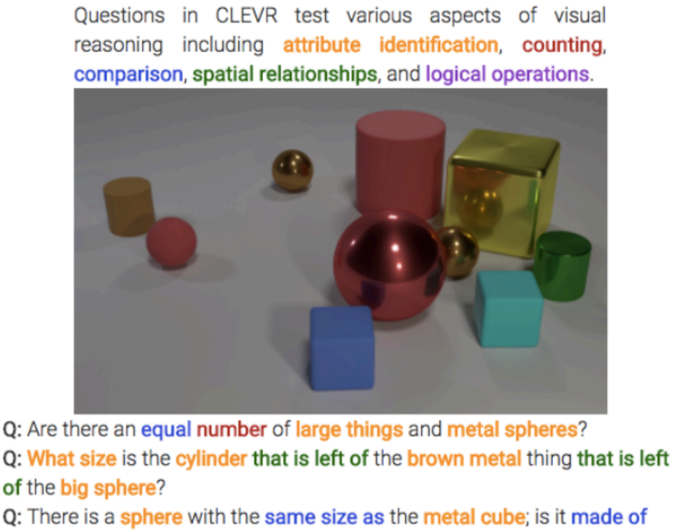

Question (Q)
What kind of book is this?
How much water can the bottle hold that is near the jacket?
What color is the backpack?
What is the black electronic device?
Why is the ground wet?
Relevant Image (I+)
Falsified Premise (P)
<book>
<bottle>
<backpack>
<device, black>
<ground, wet>
Irrelevant Image (I-)

Figure 3: **Some Examples from QRPE Dataset.** For a given question Q and a relevant image I^+, we find an irrelevant image I^- for which exactly one premise P of the question is false. If there are multiple such candidates, we select the candidate most visually most similar to I^+. As can be seen from these examples, the QRPE dataset is very challenging, with only minor visual and semantic differences separating the relevant and irrelevant images.

One caption associated with very similar images for which the sentence is T/F.

Synthetic dataset

CLEVR



Questions in CLEVR test various aspects of visual reasoning including attribute identification, counting, comparison, spatial relationships, and logical operations.

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

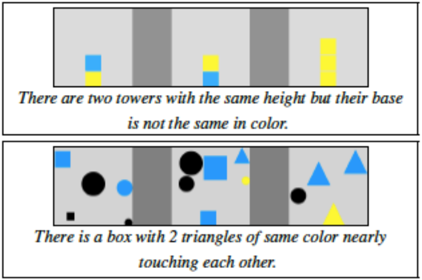
Layout

1. Cross Modal Mapping
2. Visual Phrases
3. Tasks
4. Intermezzo
5. Find their limitations
6. New task: Visual Reasoning
7. Others
8. Conclusion

Visual Reasoning

NLVR

Suhr et al. *A Corpus of Natural Language for VIsual Reasoning*. ACL 2017



There are two towers with the same height but their base is not the same in color.

There is a box with 2 triangles of same color nearly touching each other.

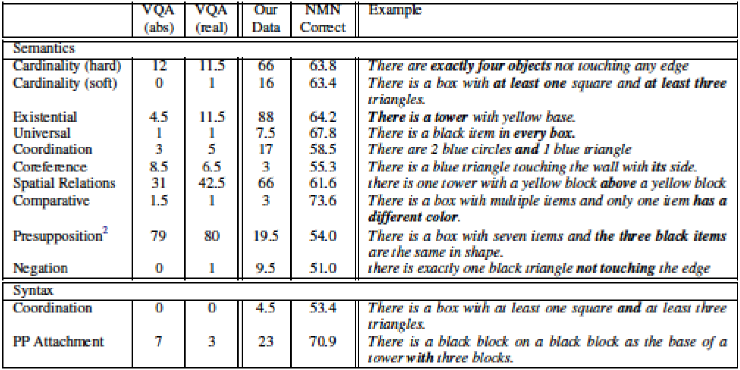
Figure 1: Example sentences and images from our corpus. Each image includes three boxes with different object types. The truth value of the top sentence is true, while the bottom is false.

Binary task: T or F.

Visual Reasoning

Results

Best performing model: Neural module network (Andreas et al 2016)



	VQA (abs)	VQA (real)	Our Data	NMN Correct	Example
Semantics					
Cardinality (hard)	12	11.5	66	63.8	*There are **exactly four objects** not touching any edge*
Cardinality (soft)	0	1	16	63.4	*There is a box with **at least one** square and **at least three** triangles.*
Existential	4.5	11.5	88	64.2	***There is a tower** with yellow base.*
Universal	1	1	7.5	67.8	*There is a black item in **every box.***
Coordination	3	5	17	58.5	*There are 2 blue circles **and** 1 blue triangle*
Coreference	8.5	6.5	3	55.3	*There is a blue triangle touching the wall with **its** side.*
Spatial Relations	31	42.5	66	61.6	*there is one tower with a yellow block **above** a yellow block*
Comparative	1.5	1	3	73.6	*There is a box with multiple items and only one item **has a different color**.*
Presupposition[2]	79	80	19.5	54.0	*There is a box with seven items and **the three black items** are the same in shape.*
Negation	0	1	9.5	51.0	*there is exactly one black triangle **not touching** the edge*
Syntax					
Coordination	0	0	4.5	53.4	*There is a box with at least one square **and** at least three triangles.*
PP Attachment	7	3	23	70.9	*There is a black block on a black block as the base of a tower **with** three blocks.*

Table 2: Qualitative and empirical analysis of our data and VQA (Antol et al., 2015). We analyze 200 sentences for each dataset. The data is categorized to semantic and syntactic categories. We use the terms *hard* and *soft* cardinality to differentiate between language using exact numerical values and ranges. For each dataset, we show the percentage of the samples analyzed that demonstrate the phenomena. We analyze abstract (abs) and real images from VQA separately. For our data, we also include the accuracy using the NMN system (Section 6) for the subset of images we tagged with this category.

Layout

1. Cross Modal Mapping
2. Visual Phrases
3. Tasks
4. Intermezzo
5. Find their limitations
6. New task: Visual Reasoning
7. Others
8. Conclusion

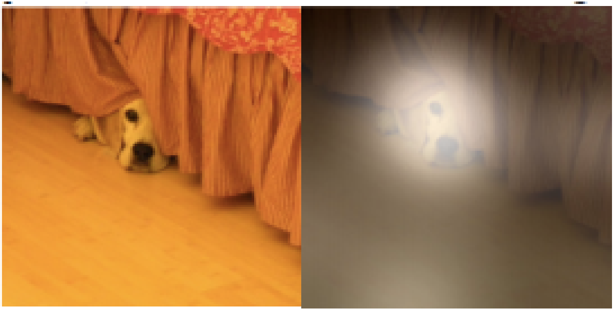
Other applications

- Spoken VQA (posted on ArXiv on the 1st of May)
- Multimodal Machine Translation
- Image Generation

Cutting-edge fancy models' ingredients

- Memory & Attention: To focus on some parts of the visual vectors (stored in the memory) e.g. by using the linguistic query to "see" the image.
- Generative adversarial networks (GAN): two neural networks competing against each other in a game framework.
- Reinforcement Learning.

Attention



A dog is standing on a hardwood floor.

Layout

1. Cross Modal Mapping
2. Visual Phrases
3. Tasks
4. Intermezzo
5. Find their limitations
6. New task: Visual Reasoning
7. Others
8. Conclusion

Conclusion

- Impressive progress
- Hard but fun to learn
- A land of new ideas can be explored

My wish:

Combine language (pragmatics) with vision.

In January we will start a reading group on Language and Vision.
Next week (Wed. 29th) talk by Sandro and Claudio (also Ravi will join us.)

Conclusion

- Impressive progress
- Hard but fun to learn
- A land of new ideas can be explored

My wish:

Combine language (pragmatics) with vision.

In January we will start a reading group on Language and Vision.
Next week (Wed. 29th) talk by Sandro and Claudio (also Ravi will join us.)

Conclusion

- Impressive progress
- Hard but fun to learn
- A land of new ideas can be explored

My wish:

Combine language (pragmatics) with vision.

In January we will start a reading group on Language and Vision.
Next week (Wed. 29th) talk by Sandro and Claudio (also Ravi will join us.)

Conclusion

- Impressive progress
- Hard but fun to learn
- A land of new ideas can be explored

My wish:

Combine language (pragmatics) with vision.

In January we will start a reading group on Language and Vision.
Next week (Wed. 29th) talk by Sandro and Claudio (also Ravi will join us.)

Other Useful Links

Neural Networks

- http://info.usherbrooke.ca/hlarochelle/neural_networks/content.html
- http://www.iro.umontreal.ca/~bengioy/dlbook/
- http://www.vlfeat.org/matconvnet/matconvnet-manual.pdf
- Blog posts: http://colah.github.io/

Other Useful Links

Language and Vision

- Describing Images in Sentences by Julia Hockenmaier `http://nlp.cs.illinois.edu/HockenmaierGroup/EACLTutorial2014/index.html`
- Vision and Language Summer Schools: 2nd edition 2016 (Malta). COST-ACTION.
- “Multimodal Learning and Reasoning”, Desmond Elliott, Douwe Kielay, and Angeliki Lazaridou (Tutorial at ACL 2016) `http://acl2016.org/index.php?article_id=59`
- Ferraro, F. and Mostafazadeh, N. and Huang, T. and Vanderwende, L. and Devlin, J. and Galley, M. and Mitchell, M. (2015). “A Survey of Current Datasets for Vision and Language Research”. Proceedings of EMNLP 2015.
- “How we teach computers to understand pictures” TED Talk by Fei Fei Li.

Language and Vision Research Groups

- Stanford Vision Lab – Le Fei Fei `http://vision.stanford.edu/`
- MIT: Antonio Torralba `http://web.mit.edu/torralba/www/`
- University of North Carolina – Tamara Berg `http://www.tamaraberg.com/`
- Virginia University – Devi Parikh `https://filebox.ece.vt.edu/~parikh/CVL.html`
- CLIC `http://clic.cimec.unitn.it/lavi/` – Us.
- Edinburgh University (M. Lapata, F. Keller)
- Facebook
- Google DeepMind
- More on the `iV&L Net` Cost Action `http://www.cost.eu/COST_Actions/ict/Actions/IC1307`