An old Artificial Intelligence dream that comes true: Merging language and vision modalities

Raffaella Bernardi

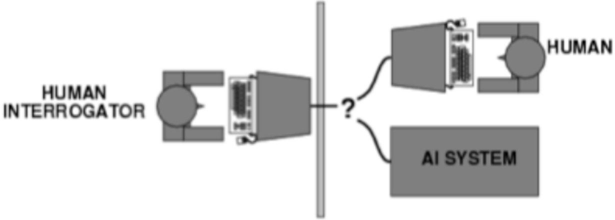
University of Trento

May 5th, 2017

Credits

L. Fei Fei, Tamara Berg, Andrej Karpathy, Angeliki Lazaridou, Elia Bruni, Marco Baroni, Chris McCormick

An old AI dream

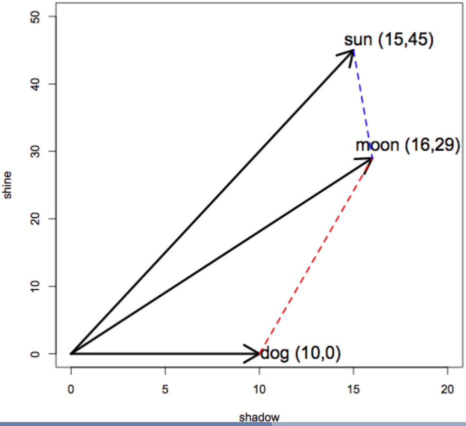


From words to Meaning Representation

Compute the word distribution

Words can be represented by vectors harvested from a corpus of texts *counting* word co-occurences.

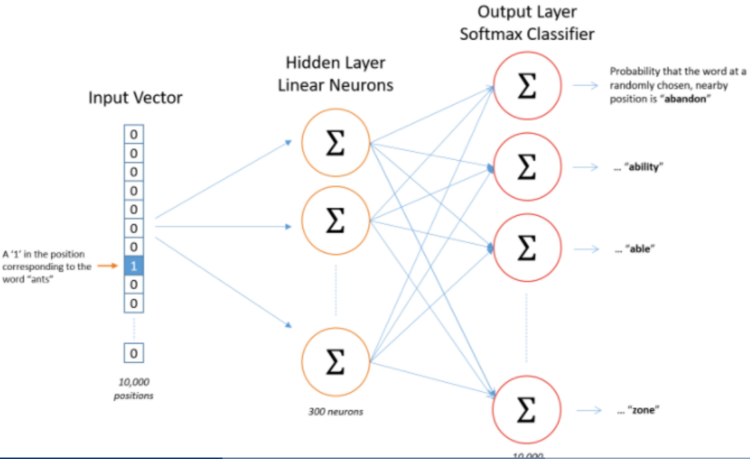
	shadow	shine
moon	16	29
sun	15	45
dog	10	0



From words to Meaning Representation

Predict the context: Word2Vec (Skip-Gram)

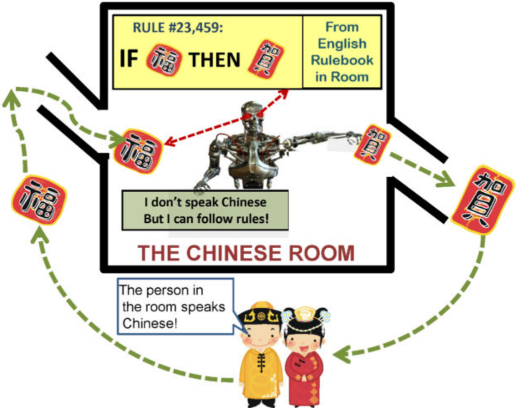
Instead counting words co-occurrences, the vector representing a word can be learned by *predicting* its nearby word. (Mikolov et al, 2013)



Vector Representations

Successful

- Lexical meaning
  - Synonyms
  - Concept categorization (eg. car ISA vehicle)
  - Selectional preferences (e.g. eat chocolate vs. *eat sympathy)
  - relation classification (exam-anxiety CAUSE-EFFECT relation)
  - salient properties (car-wheels)
- Compositionality: Phrase and Sentence
  - similairity
  - entailment



Grounding language representations into the world.

Point to the *reference* of our mental representation.

From Pixels to Meaning Representation

Gap

- To bridge the gap between pixels and “meaning”

What we see

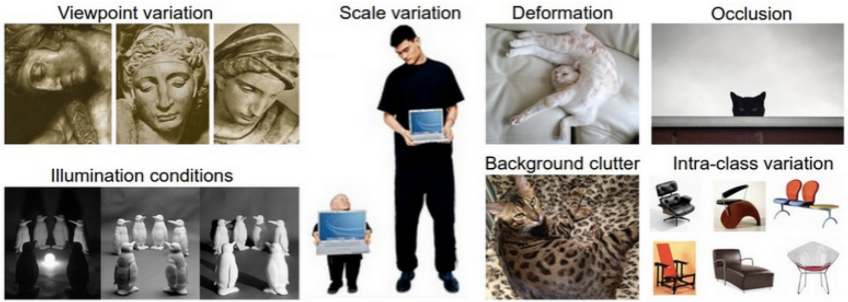
0	3	2	5	4	7	6	9	8
3	0	1	2	3	4	5	6	7
2	1	0	3	2	5	4	7	6
5	2	3	0	1	2	3	4	5
4	3	2	1	0	3	2	5	4
7	4	5	2	3	0	1	2	3
6	5	4	3	2	1	0	3	2
9	6	7	4	5	2	3	0	1
8	7	6	5	4	3	2	1	0

What a computer sees

Source: S. Narasimhan

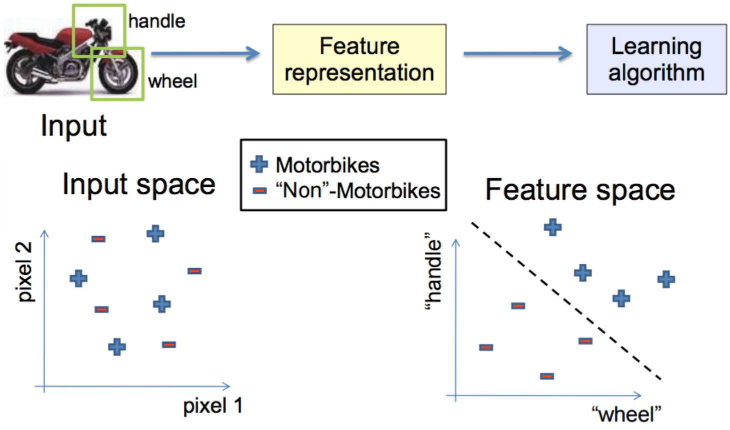
From Pixels to Meaning Representation

Challanges



From Pixels to Meaning

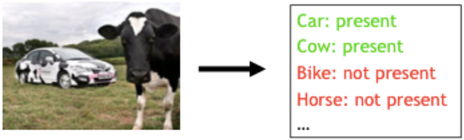
Abstract Features



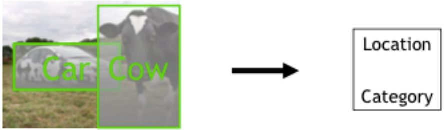
Applications: Traditional CV tasks

Objects

Image classification: assigning a label to the image.



Object localization: define the location and the category.



Similarly, scene recognition.

First Revolution: Big dataset

ImageNet

Image database organized according to the WordNet hierarchy.
Stanford Vision Lab, Stanford University & Princeton University.

- Challenges: 2007-present
- AMT: 48,940 annotators from 167 countries
- 15M images
- 22K categories of objects

From Pixels to Features

Two methods

- Bag of Visual words (BoVW) (Sivic and Zisserman, 2003)
- Convolutional neural network (CNN) (LeCun et al., 1998, Krizhevsky et ali. 2012)

From Pixels to Features: BoVW

Pipeline

Keypoints detectors To locate interesting points/content, various kinds of low-level features detectors exists:

- edge detection: the lines we would draw – encode shape info
- corner detection

Local description The identified interesting points are then described: clustered into regions and transformed into *vectors representing the region*. Several local descriptors exist, e.g:

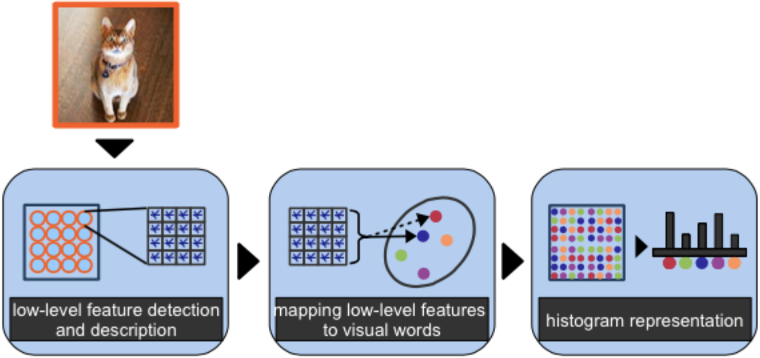
- SIFT: Scale-invariante feature transform (Lowe '99) – edge based features.
- Textons (Leung and Malik '01)
- HoG (Histograms of Oriented Gradients) (Dalal and Triggs '05)

The low-level features can capture eg. Color, Texture, Shape,

Bag of Visual Words The local descriptions are clustered to obtain the Visual Words that are used to obtain the vector representation of the image.

From Pixels to Features: BoVW

BoVW's pipeline



Second Revolution: End-to-end systems

Convolutional Neural Networks

ImageNet Classification with Deep Convolutional Neural Networks Alex Krizhevsky, Ilya Sutskever and Geoffrey E. Hinton, 2012

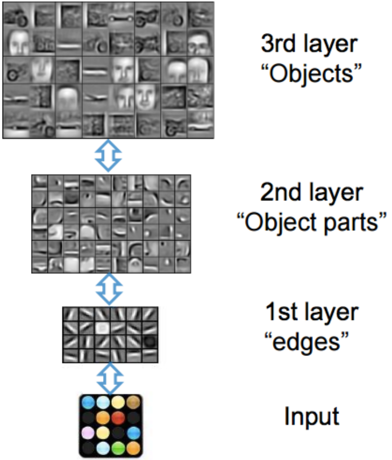
- 2012: Krizhevsky outperformed the other systems using CNN
- 2013: half of the systems used CNN
- 2014: All of the systems used CNN.

End-to-end Systems

Hierarchy of features

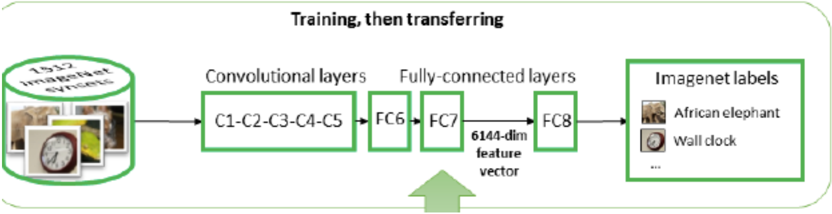
Deep Learning

- Deep architectures can be representationally efficient.
- Natural progression from low level to high level structures.
- Can share the lower-level representations for multiple tasks.



End-to-end systems

CNN: off-the-shelf vector representation

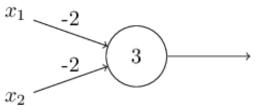


- Train a CNN on a vision task (e.g. AlexNet on ImageNet)
- Do a forward pass given an image input
- Transfer one or more layers (e.g. FC7 or C5)

Neural Networks

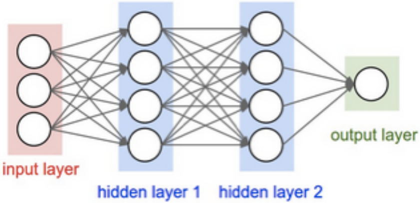
Example to compute a logical operator "Not And"

x_1	x_2	"Not and"
0	0	1
0	1	1
1	0	1
1	1	0



Input 00 produces output 1 (since: (-2)*0 + (-2) * 0 + 3 = 3 is positive) and similarly, 01 and 10; but the input 11 produces output 0 (since: (-2)*1 + (-2) * 1 + 3 = -1 is negative.)

It's a composition of functions (neurons) that goes from an n-dimensional vector to class scores.



Each neuron receives some inputs, performs a dot product and optionally follows it with a non-linearity. On the last (fully-connected) layer, they have a loss function (e.g., Softmax).

Traditional neural networks cannot use the information "about previous inputs" to inform later ones.

- *Recurrent neural networks* (RNNs) address this issue: They are networks with loops in them, allowing information to persist. They work well with short dependencies.
- *Long Short Term Memory* (LSTM) are a special kind of RNN, capable of learning long-term dependencies.

Language and Vision

Language and Visual Space can be combined!

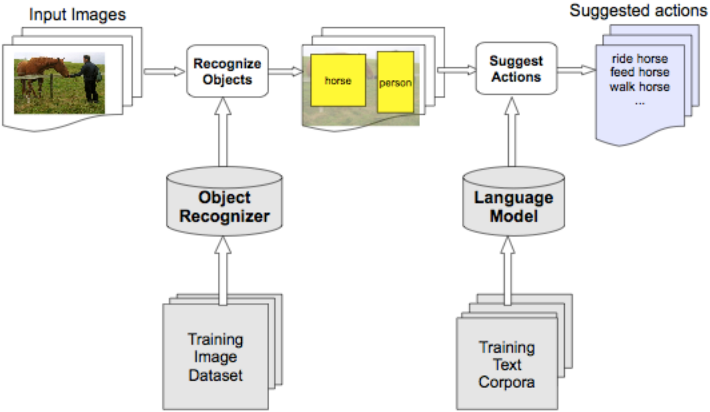
A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, S. Belongie (ICCV 2007) Objects in Context.

Not a Lemon, it's more probable a Tennis Ball. Info come from a KB (word similarity list, exctracted from internet – Google Sets).

Applications: Traditional CV tasks

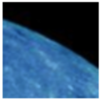
Corpora as KB source: Action recognition

Thu Le Dieu, Jasper Uijlings and R. Bernardi (2010, 2011)



Applications: Traditional NLP tasks

E. Bruni, G.B. Tran and M. Baroni (GEMS 2011, ACL 2012, Journal of AI 2014), E. Bruni, G. Boleda, M. Baroni and N. Tran (ACL 2012)

	planet	night		
moon	10	22	22	0
sun	14	10	15	0
dog	0	4	0	20

Applications: Traditional NLP tasks

Task 1 Predicting human **semantic relatedness** judgments

Improved!

Task 2 **Concept categorization**, i.e. grouping words into classes based on their semantic relatedness (*car* ISA *vehicle*; *banana* ISA *fruit*)

Improved!

Task 3 Find **typical color** of concrete objects (**cardboard is brown**, **tomato is red**)

Improved!

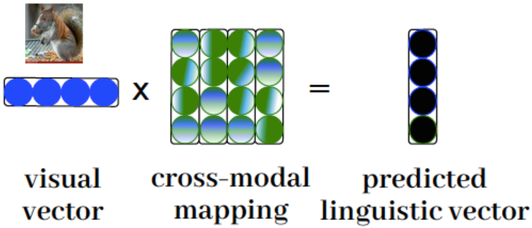
Task 4 Distinguish **literal vs. non-literal** usages of color adjectives (**blue uniform** vs. **blue note**)

Improved!

New Language and Vision Tasks

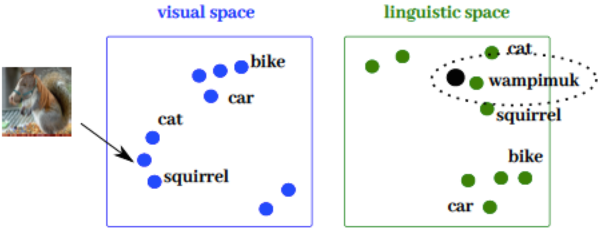
Zero-shot (or Cross-modal) Mapping: training

"Is this a wampimuk? Cross-modal mapping between distributional semantics and the visual world" A. Lazaridou, E. Bruni and M. Baroni (ACL 2015)



New Language and Vision Tasks

Zero-shot (or Cross-modal) Mapping: testing



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**

New Language and Vision tasks

Fast Mapping

New LaVi Applications: Image Captioning

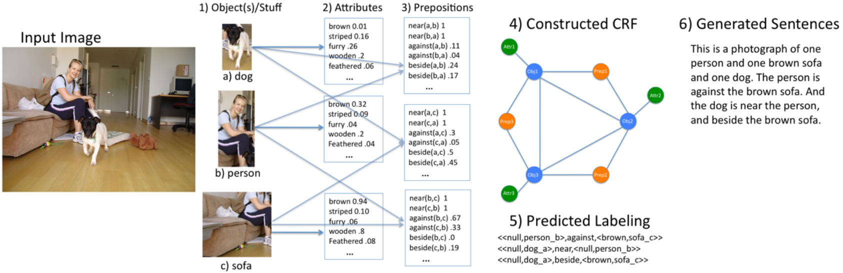
a man is throwing a frisbee in a park

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

New LaVi Applications: IC

Approach: Pipeline

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



New LaVi Applications: IC

Approach: End-to-end

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

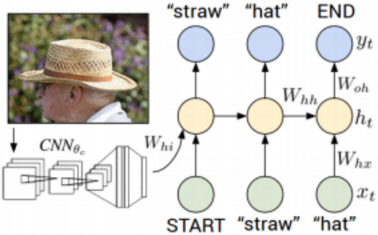


Figure 4. Diagram of our multimodal Recurrent Neural Network generative model. The RNN takes a word, the context from previous time steps and defines a distribution over the next word in the sentence. The RNN is conditioned on the image information at the first time step. START and END are special tokens.

New LaVi Applications: IC

Further info

- **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`

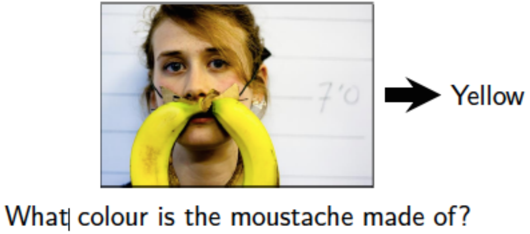
New LaVi Applications: IC

Limitations

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

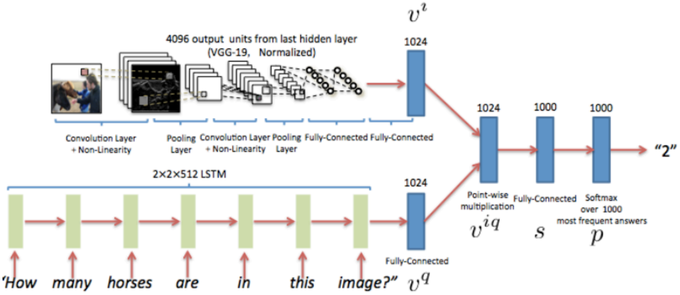
New LaVi Applications: VQA

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



New LaVi Applications: VQA

Model



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

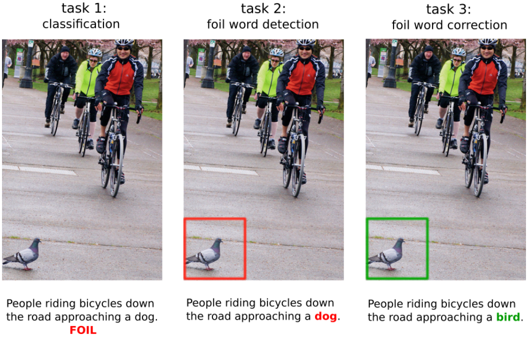
New LaVi Applications: VQA

Limitations

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

New LaVi Applications: VQA

Similar images different answers



New LaVi Applications: VQA

Further info

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

Other applications

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

Can we (linguists) be happy?

Your answer!

New LaVi Applications: IVQA

Interactive Visual Question Answering

de Vries et ali. “GuessWhat?! Visual object discovery throught multi-modal dialogue” (2017 arXiv)
Goal of the game: locate an unknown object in a rich image scene by asking a sequence of questions.

Questioner	Oracle
Is it a vase?	Yes
Is it partially visible?	No
Is it in the left corner?	No
Is it the turquoise and purple one?	Yes

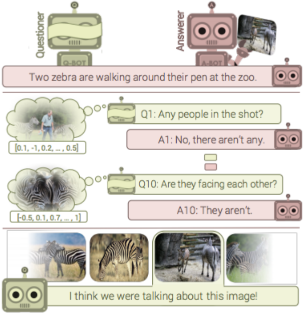
Figure 1: An example game. After a sequence of four questions, it becomes possible to locate the object (highlighted by a green bounding box).

Training data: human-playes games: 800K visual QA pairs on 66K images.

New LaVi Applications: IVQA

Interactive Visual Question Answering

Das et ali. “Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning” (2017 arXiv)



Cooperative guessing game: Q-Bot has to select a particular unseen image among several ones that the A-Bot sees. The two agents communicate through NL. Trained on VisDial dataset.

Cutting-edge fancy models' ingredients

- Generative adversarial networks (GAN): two neural networks competing against each other in a game framework.
- 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.

Attention

A dog is standing on a hardwood floor.

Attention

Quantifiers

Pay attention to those sets! Learning quantification from images Sorodoc et. al. (Submitted)

(a) 
(b) 
(c) 
(d) 
(e) 

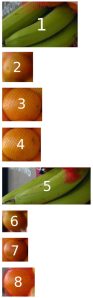
Query: ___ fish are red.

Answers: (a) All, (b) Most, (c) Some, (d) Few, (e) No.

Datasets: Q-COCO

ORIGINAL IMAGE

SCENARIO



ANNOTATION

banana: healthy

orange: fresh, tasty/delicious

orange: healthy, tasty/delicious, appetizing, fresh, round

orange: tasty/delicious, appetizing, fresh, cooked

banana: laying, healthy, tasty/delicious, horizontal, fresh, whole

orange: laying, round, fresh, appetizing tasty/delicious, whole, healthy

orange: tasty/delicious, fresh

orange: fresh

GENERATED QUERIES	PROPORTION	GROUND-TRUTH ANSWER
1. ____ oranges are fresh	100%	**all**
2. ____ oranges are whole	16.7%	**few**
3. ____ oranges are healthy	33.3%	**some**
4. ____ oranges are tasty/delicious	83.3%	**most**
5. ____ oranges are horizontal	0%	**no**

Datasets: Q-ImageNet

ORIGINAL IMAGE	SCENARIO	ANNOTATION

dog: furry, black

dog: furry, black

dog: furry, black, smooth

rabbit: furry, white, brown

dog: furry, black, brown, smooth

dog: furry, black, gray

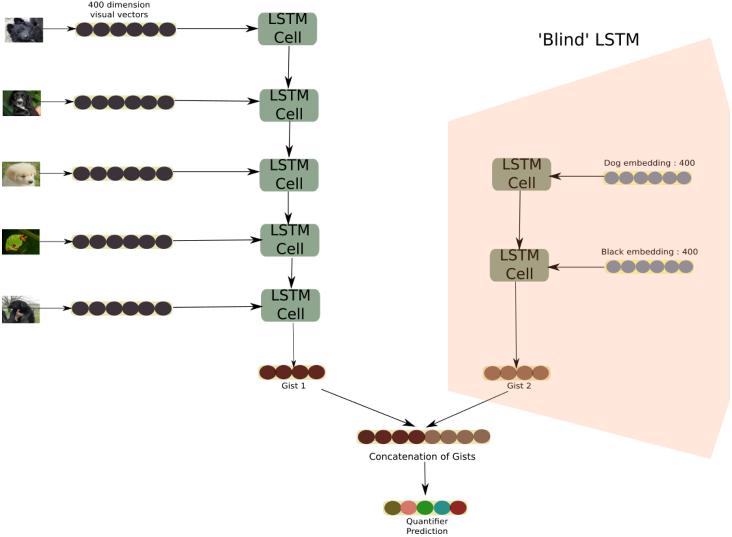
hoop: white, red, round

dog: black, white

GENERATED QUERIES	PROPORTION	GROUND-TRUTH ANSWER
1. ____ dogs are black	100%	**all**
2. ____ dogs are white	16.7%	**few**
3. ____ dogs are smooth	33.3%	**some**
4. ____ dogs are furry	83.3%	**most**
5. ____ dogs are red	0%	**no**

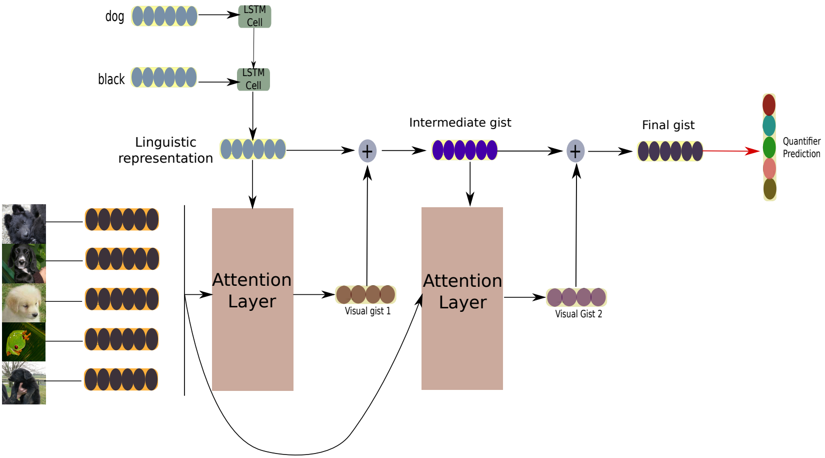
Sequential Processing

CNN+LSTM model



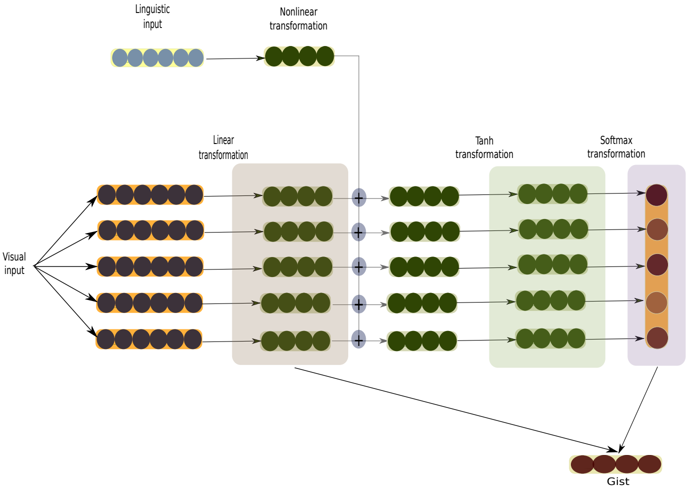
Stacked Attention Model

Yang, Z., et al. (CVPR 2016). Stacked attention networks for image question answering.

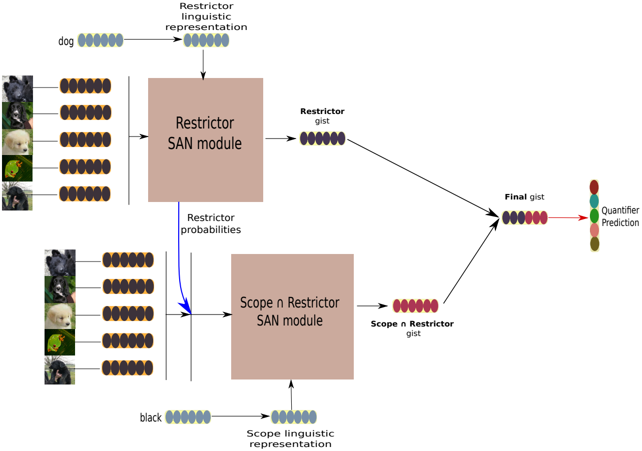


Attention Mechanism: SAN's attention layer

Yang, Z., et al. (CVPR 2016). Stacked attention networks (SAN) for image question answering.



Linguistically motivated NNs with stacked attention



Conclusion

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

My wish:

Combine language (pragmatics) with vision.

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`

The team@UniTN

Ionut

Sandro

Ravi

Aurelie

me