

Language and Vision

Intro

Raffaella Bernardi
University of Trento



UNIVERSITÀ DEGLI STUDI
DI TRENTO

Distributional Semantics: A successful story..

Lexical meaning

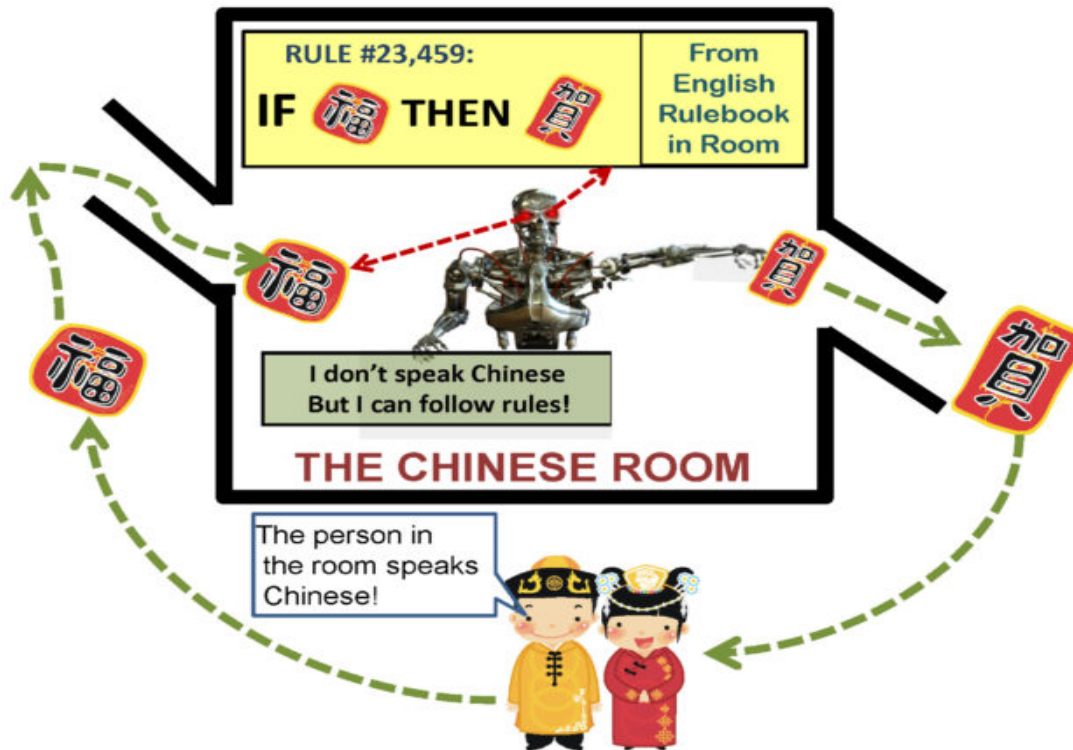
- Synonyms
- Concept categorization (eg. car ISA vehicle)
- Selectional preferences (e.g. eat chocolate vs. *eat sympathy)
- Relation classification (exam-anxiety CAUSE-EFFECT relation)
- Salient properties (car-wheels)

Compositionality: Phrase and Sentence

- Similarity
- Entailment

Distributional Semantics: .. but Grounding Problem

Grounding language representation into the world:
point to the reference of our mental representation.



Computer Vision: From pixels to Meaning

- 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

How to represent an image?



Compute average pixel

By regions

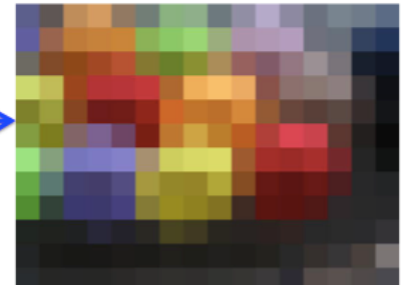


Photo by: [marilite](#)

Image representation challenge: Invariance

Viewpoint variation



Scale variation



Deformation



Occlusion



Illumination conditions



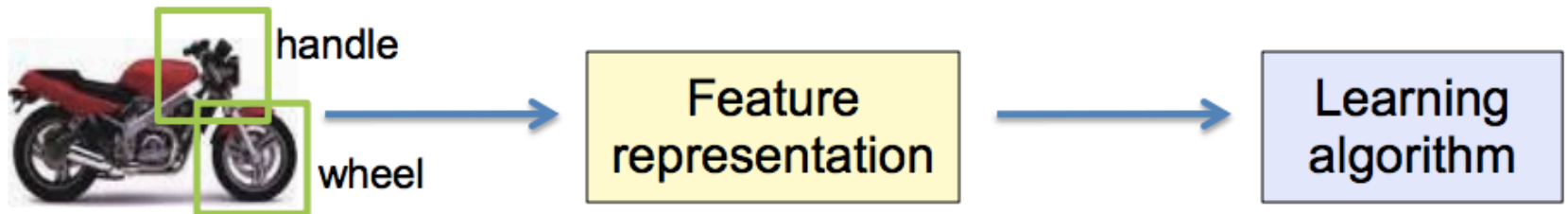
Background clutter



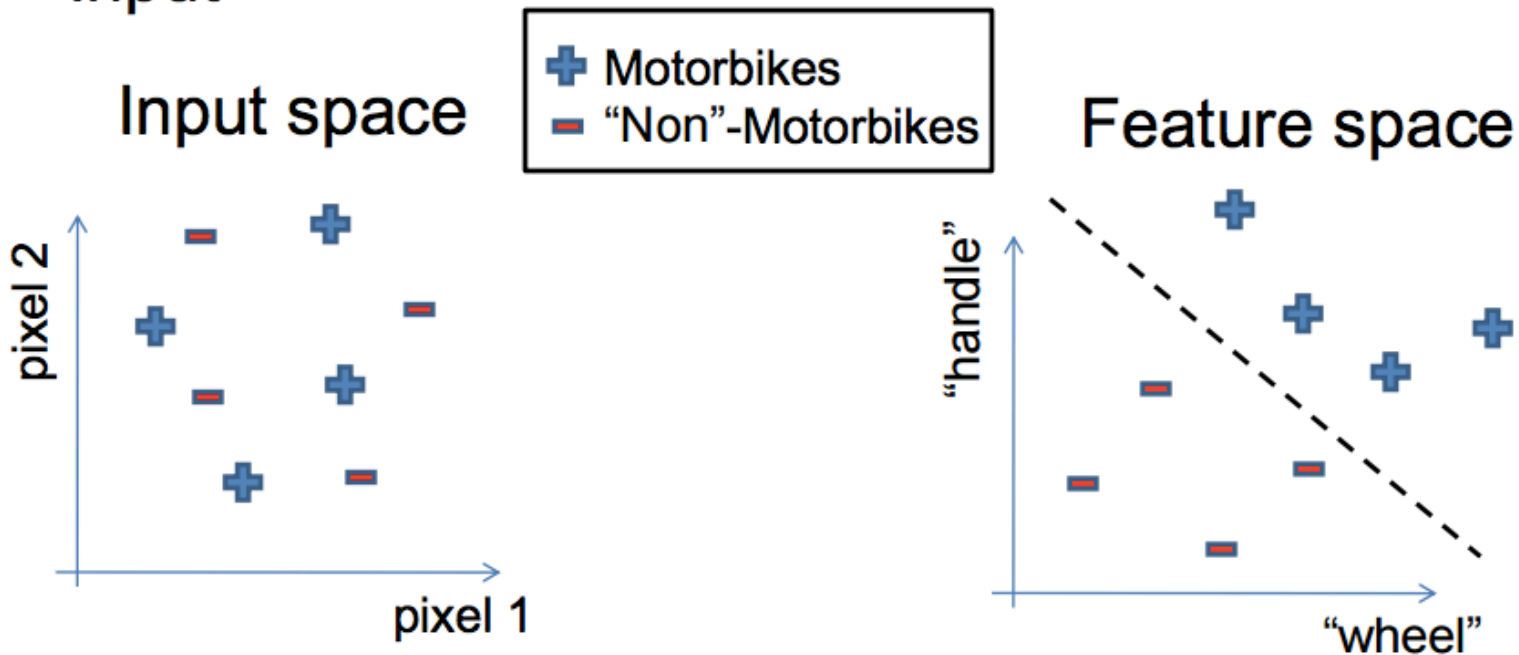
Intra-class variation



CV long standing challenge: Abstract Features



Input



CV traditional tasks: Objects

Image classification:



Car: present
Cow: present
Bike: not present
Horse: not present
...

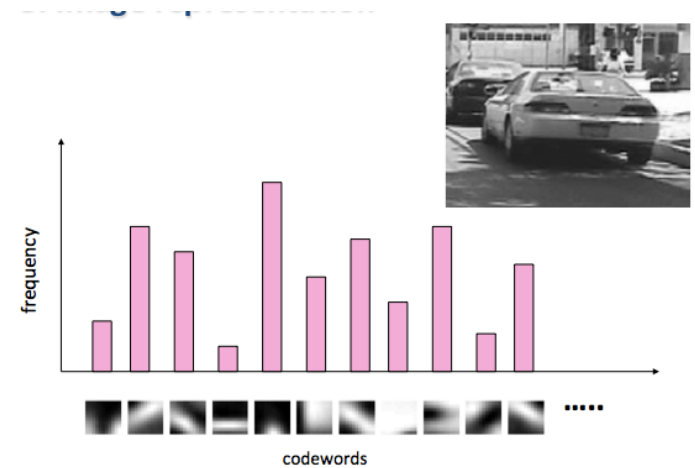
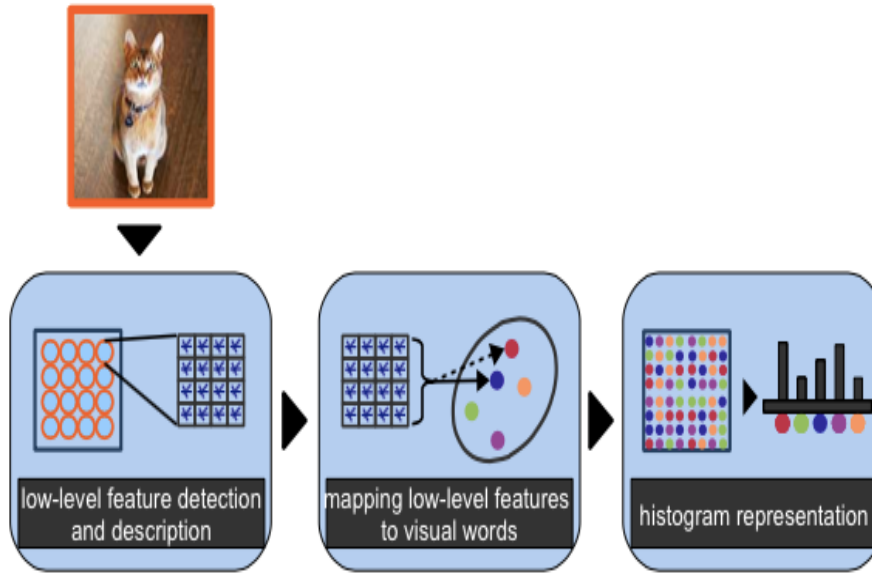
Object localization:



Location
Category

From objects to scene classification

Traditional Method: Pipeline



CV first important revolution: ImageNet

ImageNet:

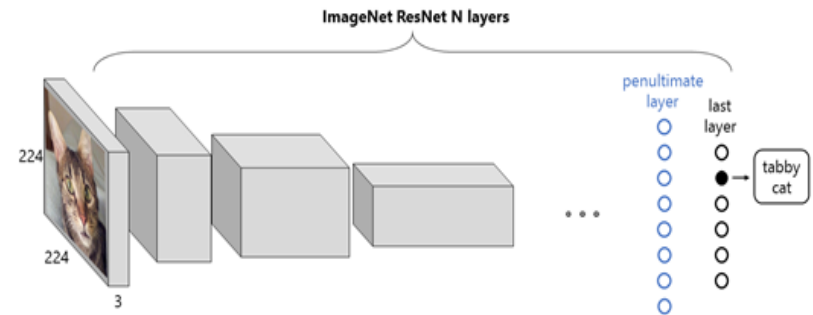
- Stanford Vision Lab, Stanford University & Princeton University.
- Image database organized according to the WordNet hierarchy.
- Challenges: 2007-present
- AMT: 48,940 annotators from 167 countries
- 15M images
- 22K categories of objects

CV second important revolution: Convolutional Neural Networks

ImageNet Classification with Deep Convolutional Neural Networks

*Alex Krizhevsky, Ilya Sutskever
and Georey 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.



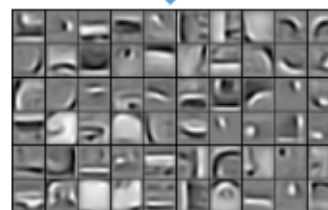
CNN: Hierarchy of features

Deep Learning

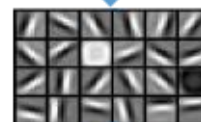
- Deep architectures can be representationally efficient.
- Natural progression from low level to high level structures.
- Can share the lower-level representations for multiple tasks.



3rd layer
“Objects”



2nd layer
“Object parts”

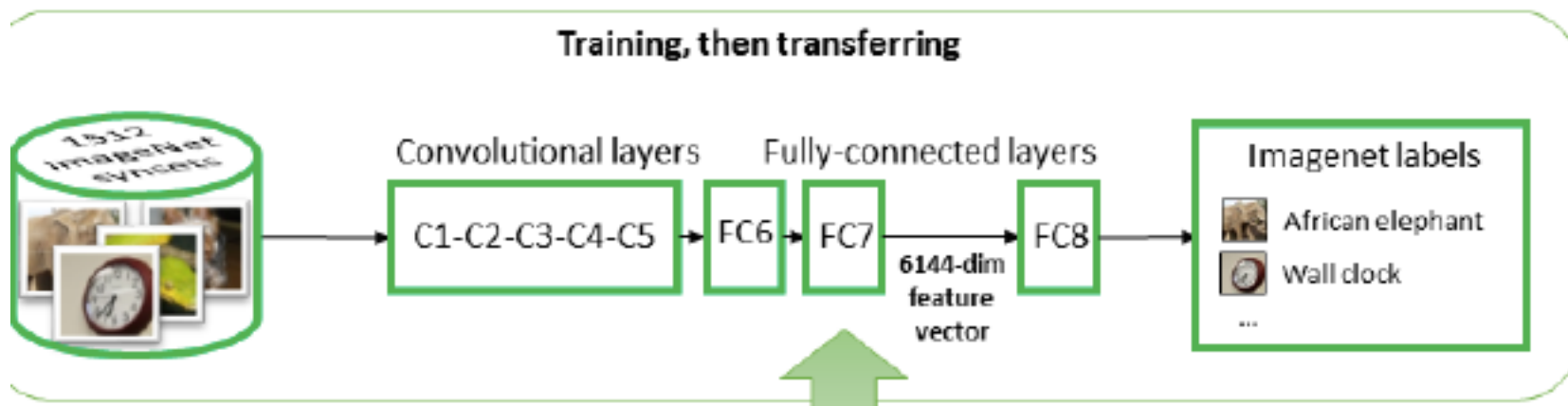


1st layer
“edges”



Input

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)

Language and Vision

Language and Visual Spaces can be combined!

Cognitive Angle:

Language and Vision Representations

must be combined!

Applied Angle:

Combining Language and Vision Representations

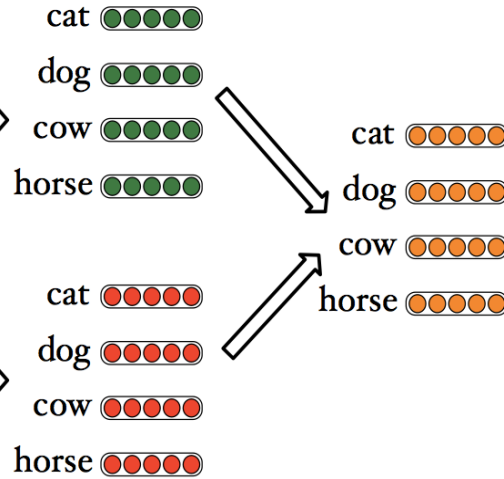
is **very useful**

Language and Vision

- Multimodal Representations:
 - learned separately and concatenated
 - learned separately and translated one into the other
 - learned jointly
- Multimodal Tasks:
 - Exploit language to improve on traditional CV tasks
 - Exploit vision to improve on traditional NLP tasks
 - New Multimodal Tasks

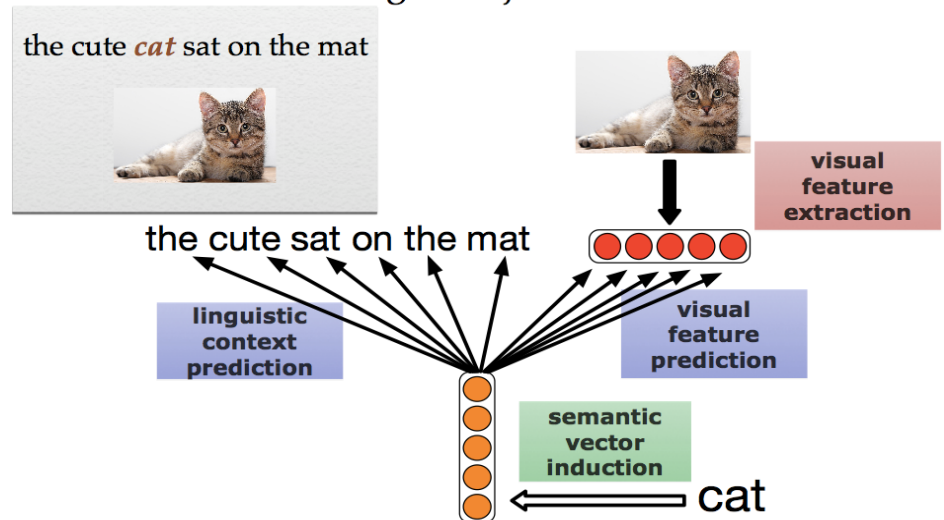
Multimodal Representation

Lucifer Sam, siam cat. Always sitting by your side
Always by your side. That cat's something I can't
explain. Ginger, ginger, Jennifer Gentle you're a witch.
You're the left side He's the right side. Oh, no! That
cat's something I can't explain. Lucifer go to sea. Be a
hip cat, be a ship's cat. Somewhere, anywhere. That
cat's something I can't explain. At night prowling sifting
sand. Hiding around on the ground. He'll be found
when you're around. That cat's something I can't explain



Multimodal Distributional Semantics
Bruni, Tran and Baroni (2014)

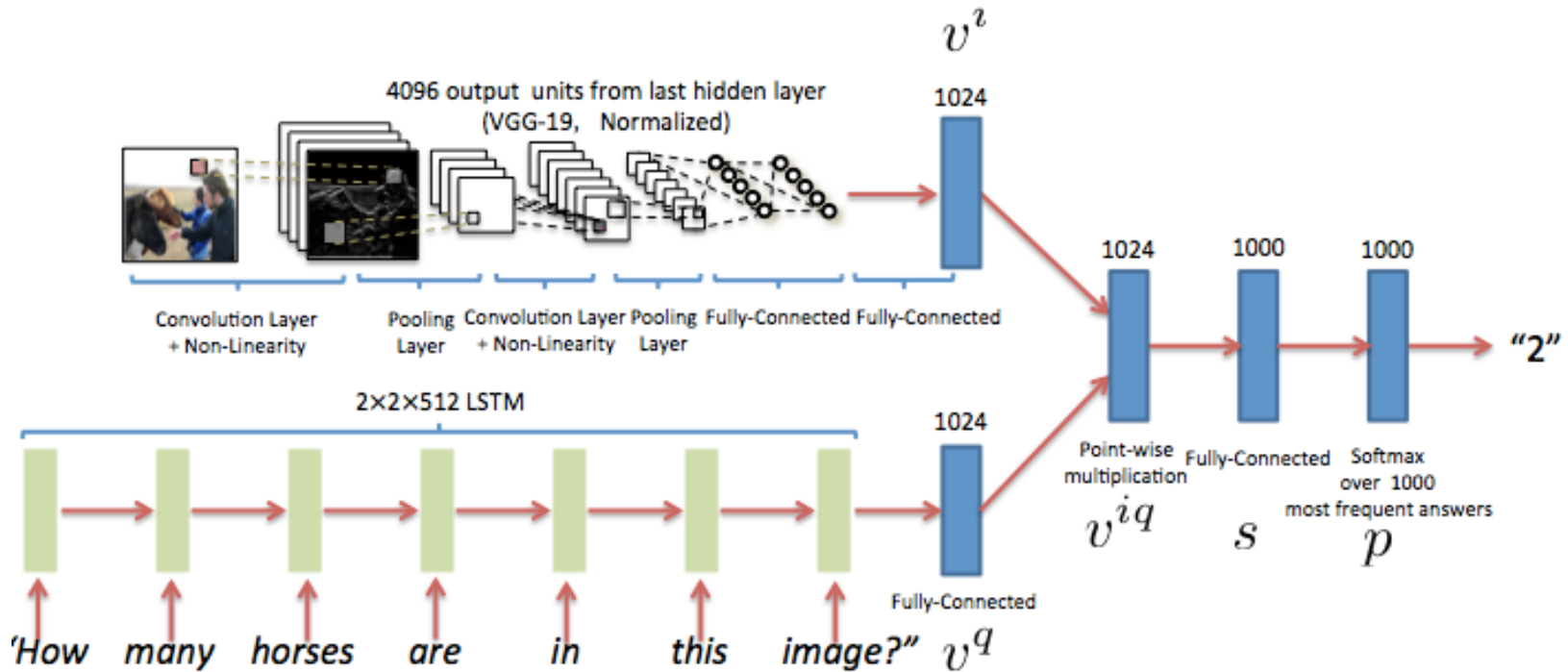
Learning from joint contexts



*Combining Language and Vision with a
Multimodal Skipgram Model*

Lazaridou, Phan and Baroni (2015)

Basic Multimodal Models: Point-wise multiplication



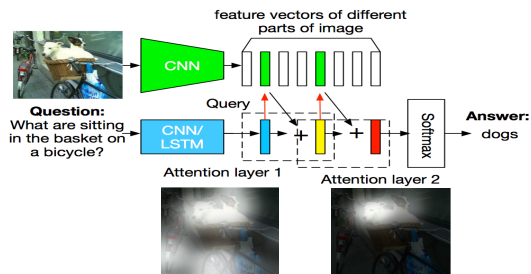
$$v^{iq} = v^i \circ v^q$$

$$s = Wv^{iq} + b$$

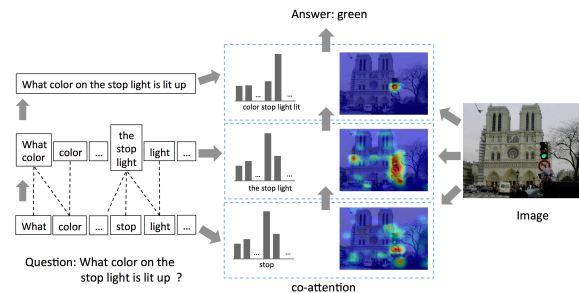
$$p_a = \frac{e^{s_a}}{\sum_{a'} e^{s_{a'}}$$

What has the community gained?

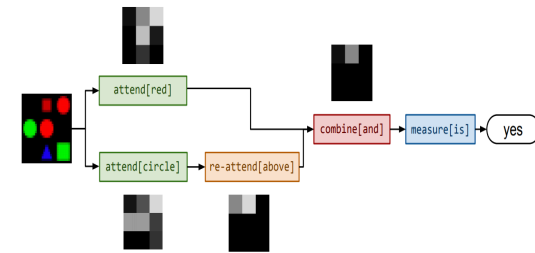
- Attention Networks
- Hierarchical Co-attention
- Bottom-up Top-down attention
- Compositionality
- Multi-modal Pooling



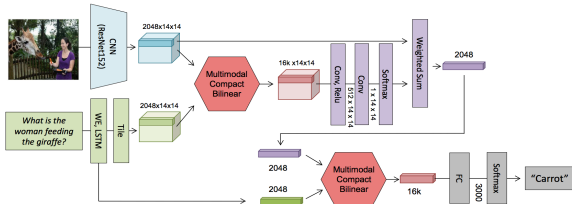
Stacked Attention Networks
Yang et al., CVPR 16



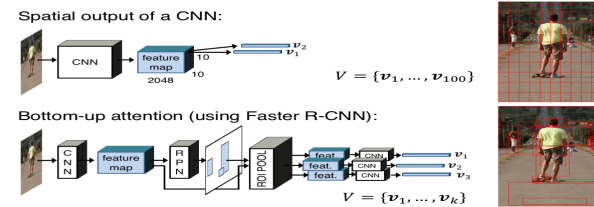
Hierarchical Question-Image Co-Attention
Lu et al., NIPS 16



Neural Module Networks
Andreas et al., CVPR 16



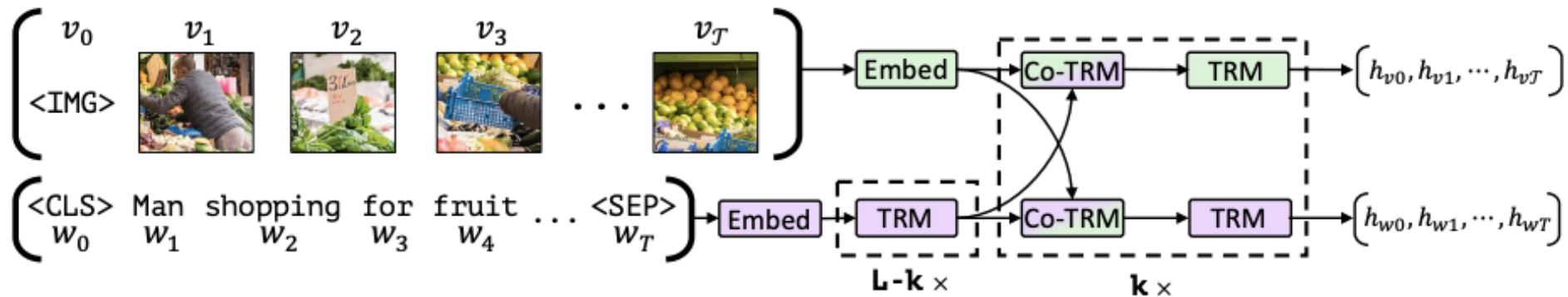
Multimodal Compact Bilinear Pooling
Fukui et al., EMNLP 16



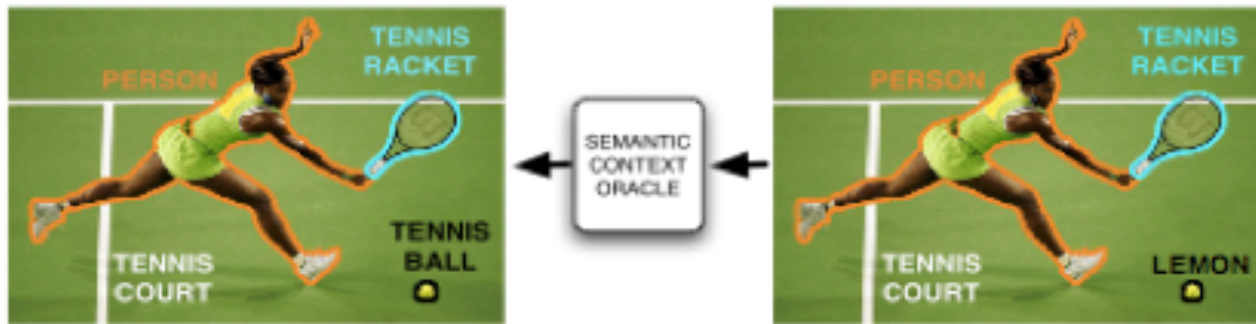
Bottom-Up and Top-Down Attention
Anderson et al., CVPR 18

Task Agnostic Representations

Vision-Language Bert



Multimodal Tasks: Improve traditional CV tasks

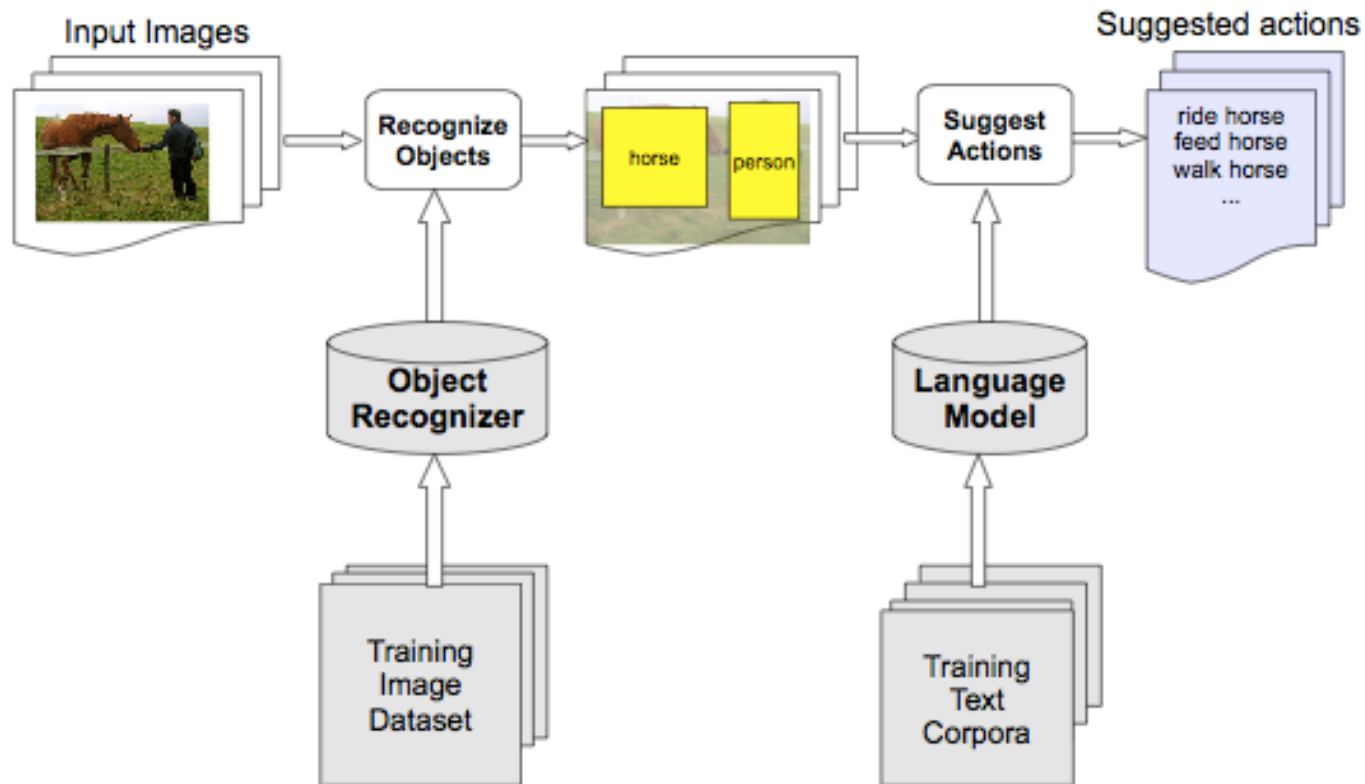


Not a lemon, it's more probable a tennis ball. -- Info come from a KB (word similarity list, extracted from internet Google Sets).

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

Multimodal Tasks:

Improve CV task: Action recognition



Use of Corpora for Action Recognition.

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

Multimodal Tasks:

Improve 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)

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!

Multimodal tasks: Lexical Preferences

Is the argument
migas a
plausible object
for the
predicate *eat*?

Can you *eat*
“*migas*”?



Can you *eat*
“*carillon*”?



Can you *eat*
“*mamey*”?



5

Multimodal Tasks: Machine Translation

Images for "candle" (English)



Images for "vela" (Spanish)

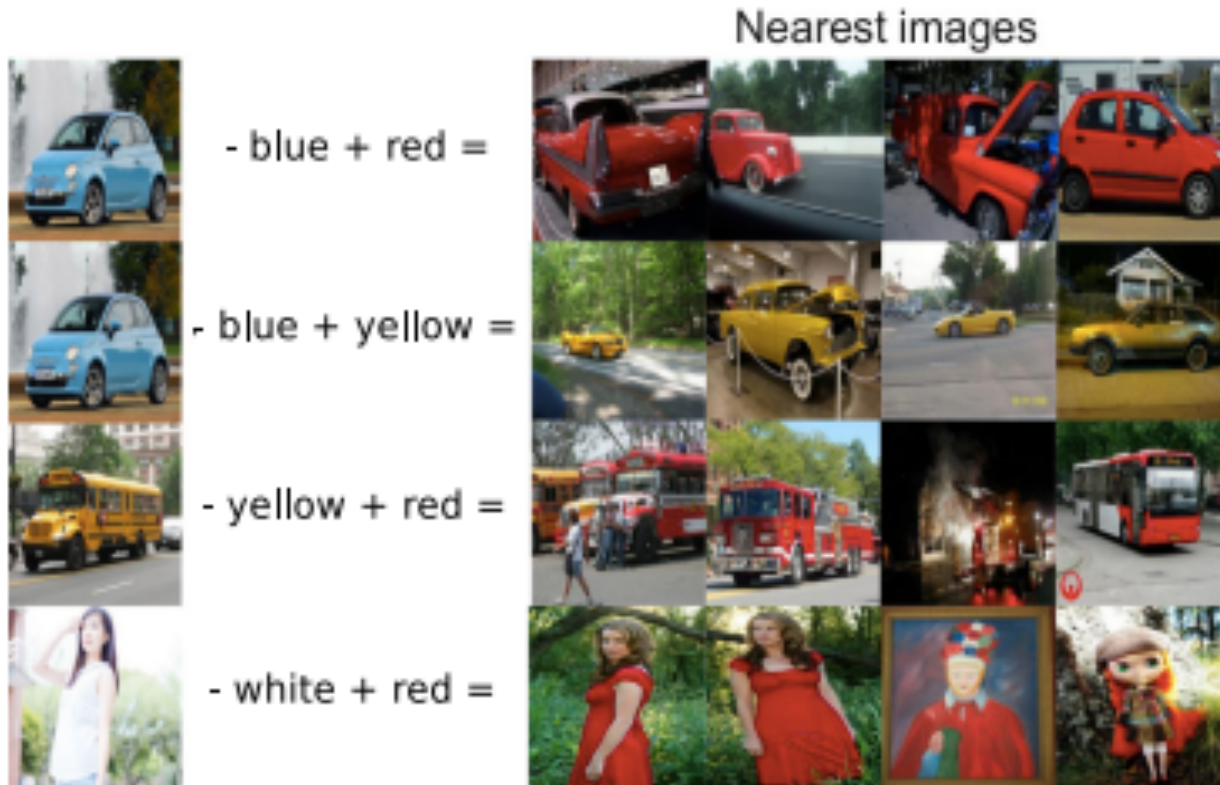
Figure 1: Matching words through their images: Images retrieved from the web for the English word *candle* (top) and the Spanish word *vela* (bottom). The matching between detected SIFT keypoints is shown for a pair of images.

S. Bergsma, B. Van Durme, (IJCAI 2011)

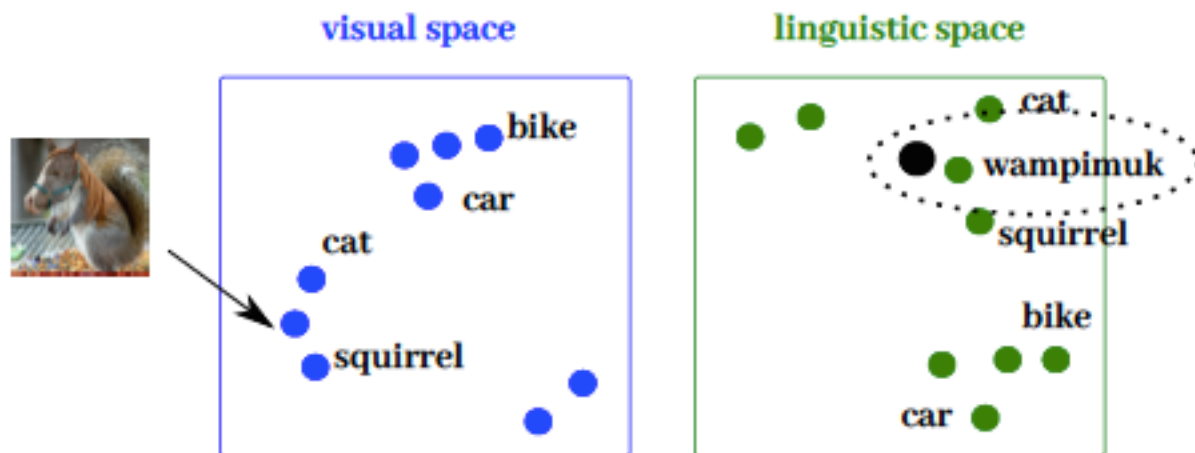
Learning Bilingual Lexicons using the Visual Similarity of Labeled Web Images,

For recent work look for eg. Lucia Specia,

Multimodal Vector Spaces

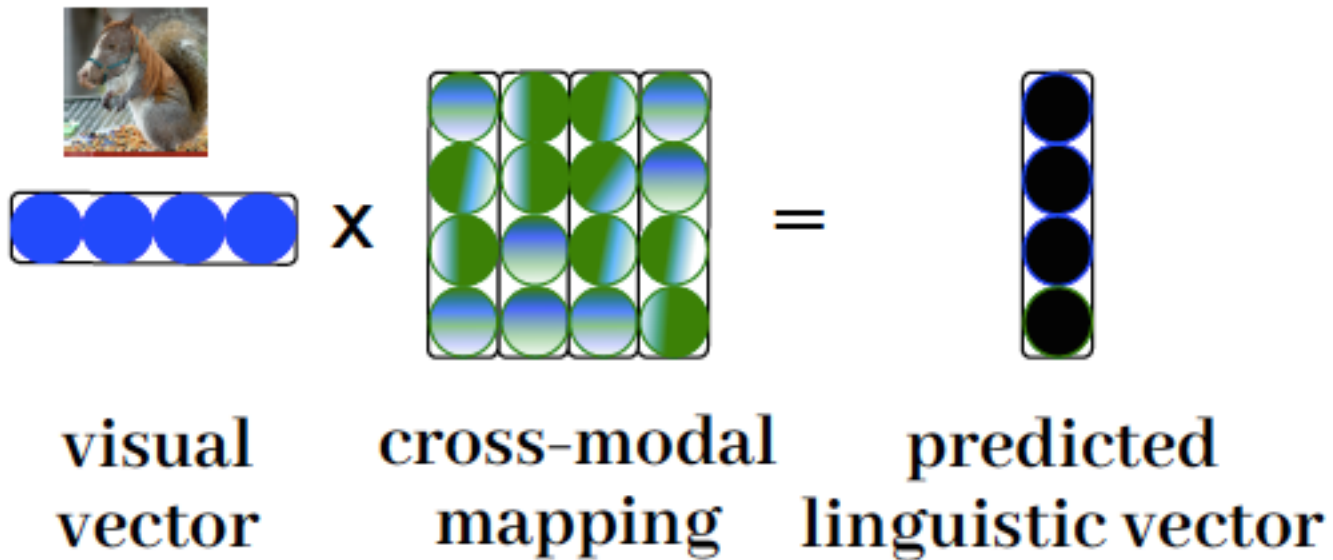


New Multimodal Tasks: Cross-Modal Mapping



- 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 Multimodal Tasks: Zero-shot learning



New Multimodal Tasks: Fast Mapping

New concepts are assumed to be encountered in a limited linguistic context and therefore lacking rich linguistic representations.

Lazaridou et al NAACL 2016 *Multimodal semantic learning from child-directed input*

New Multimodal Tasks: Composition

- Assumption from linguistics: Adjectives are noun modifiers. They are functions from N into N.
- Lexical Function model:: adjectives can be learned from (ADJ N, N) inputs.
- Applied to images: Compositional Visual Model?

New Multimodal Tasks: Image Captioning (IC)



a man is throwing a frisbee in a park

- **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 \(2015\)](#):

Limitations:

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

New Multimodal Tasks: Visual Question Answering (VQA)



➔ Yellow

What colour is the moustache made of?

Datasets: DAQUAR 2014, COCO-QA, VQA, Visual7W, Visual Genome, VisWiz

Survey: Visual Question Answering: A Survey of Methods and Datasets Wu et al, (2016)

Limitations:

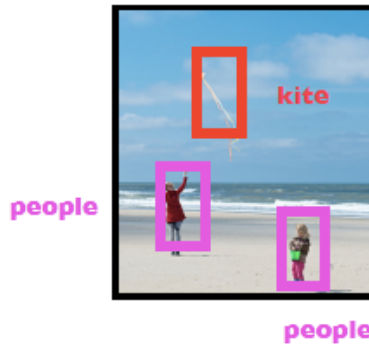
- Language prior problem: Blind models perform pretty well (50% accuracy on COCO-VQA!).
- ➔ But see development of new real image datasets: VQA2, TDIUC

New Multimodal Tasks

Datasets: Faces in the Wild, Flickr 30k Entities, VRD, Visual Genome

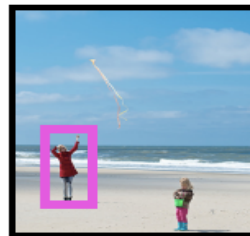
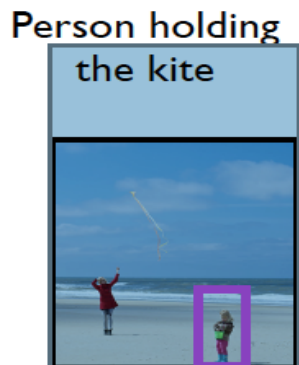
Duygulu et al 2002, Barnard et al 2003, Berg et al 2004, Plummer et al 2015, Karpathy and Fei-Fei 2015, Zhu et al 2015, Krishna et al 2016, Lu et al 2016

Image-Text Alignment



Two **people** playing with a **kite** on the beach

Referring Expressions



kid staring at the kite

Datasets: D-TUNA Corpus, Referit Game Dataset, Referit Game MS-COCO

Mitchell et al 2013, Fitzgerald et al 2013, Kazemzadeh et al 2014, Mao et al 2015, Yu et al 2016, Hu et al 2016, Yu et al 2017, Nagaraja et al 2016, Fang et al 2015

Other even more recent

New Multimodal Tasks:

- Spoken VQA
- Multimodal Machine Translation
- Image Generation
- Visual Dialogue

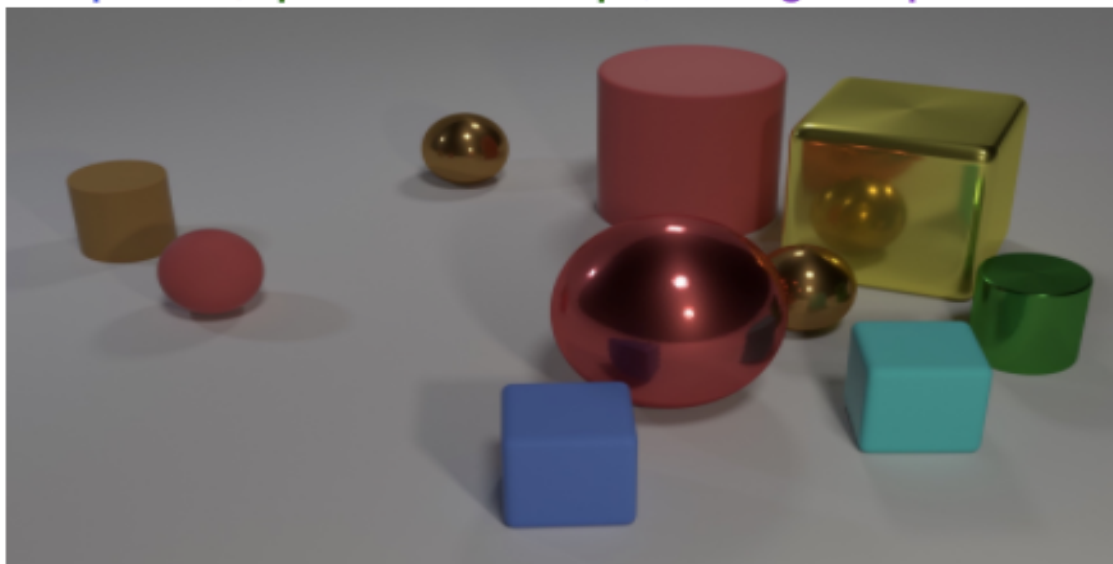
- Visual Story Telling (Huang et al. 2016)
- Question Generation (Mostafazadeh et al 2016, Jain et al 2017)
- Explanation (Park et al. 2018), Counter-factual (Hendricks et al. 2018), Inferences (Iyyer et al. 2017), Entailment (Vu et al. 2018)
- Emotion recognition, You et al. 2016
- Learning to quantify (vague quantifiers, exact numbers). Pezzelle et al. 2016, 2017, 2018
-

Cutting-edge fancy models: Learning Paradigms

- Adversarial learning
- Reinforcement Learning
- Cooperative Learning
- ...

New Multimodal Tasks Diagnostic 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 the same material** as the **small red sphere**?

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

Jonhson et al CVRP 2017: <https://cs.stanford.edu/people/jcjohns/clevr/>

New Multimodal Tasks: Diagnostic Dataset: NLVR

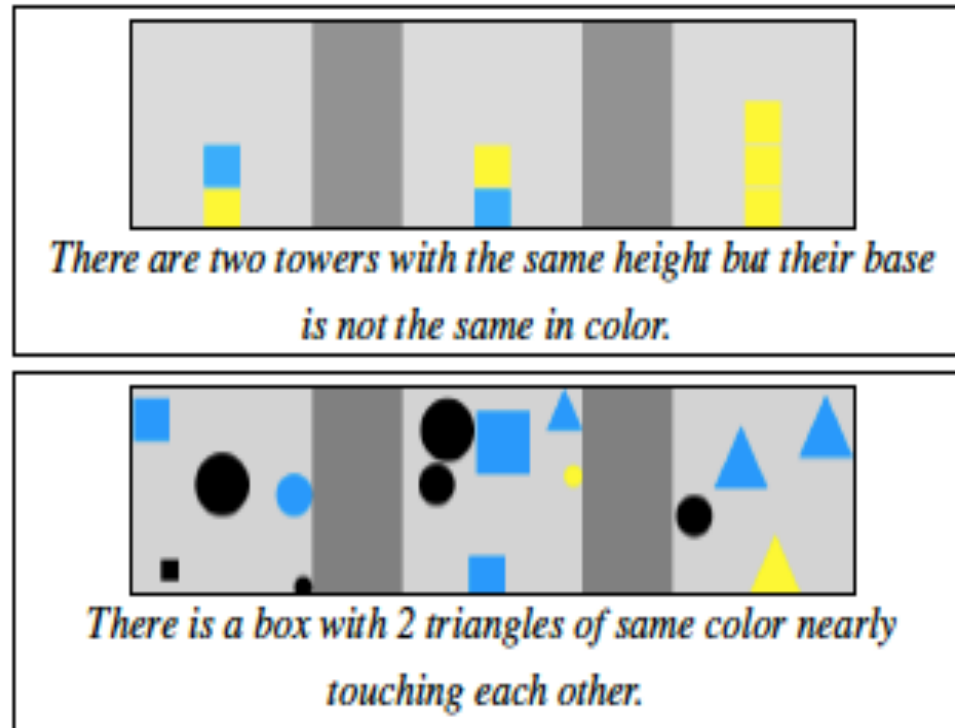


Figure 1: Example sentences and images from our corpus. Each image includes three boxes with different object types. The truth value of the top sentence is true, while the bottom is false.

New Multimodal Tasks

Diagnostic Dataset: 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**.

Importance of Understanding NN models

[SiVL](#)@ECCV Shortcomings in Vision and
Language

[BlackboxNLP](#)@EMNLP

Surveys/Tutorials

- ACL 2017:
<https://www.cs.cmu.edu/~morency/MMML-Tutorial-ACL2017.pdf>
- COLING 2018
<https://arxiv.org/abs/1806.06371>
- Baroni 2015: *Grounding DS into the Visual Worlds*
- Kafle et al. 2019: Survey

Some 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/>
- Edinburgh University (M. Lapata, F. Keller)
- University of Sheffield Lucia Specia <http://staffwww.dcs.shef.ac.uk/people/L.Specia/>
- Universitat Pompeu Fabra, COLT group, Gemma Boleda: <http://gboleda.utcompling.com/>
- University of Amsterdam, Raquel Fernandez

- Facebook FAIR
- Google DeepMind
- More on the iV&L Net Cost Action http://www.cost.eu/COST_Actions/ict/Actions/IC1307

References on tasks

- Jain, U., Zhang, Z., Schwing, A.G.: Creativity: Generating Diverse Questions using Variational Autoencoders. In: CVPR. (2017) 5415-5424
 - Li, Y., Huang, C., Tang, X., Change Loy, C.: Learning to Disambiguate by Asking Discriminative Questions. In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 3419-3428
 - Vondrick, C., Oktay, D., Pirsiavash, H., Torralba, A.: Predicting motivations of actions by leveraging text. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 2997-3005
 - Park, D.H., Hendricks, L.A., Akata, Z., Rohrbach, A., Schiele, B., Darrell, T., Rohrbach, M.: Multimodal Explanations: Justifying Decisions and Pointing to the Evidence. In: 31st IEEE Conference on Computer Vision and Pattern Recognition. (2018)
 - Hendricks, L.A., Hu, R., Darrell, T., Akata, Z.: Generating Counterfactual Explanations with Natural Language. arXiv preprint arXiv:1806.09809 (2018)
 - You, Q., Luo, J., Jin, H., Yang, J.: Building a Large Scale Dataset for Image Emotion Recognition: The Fine Print and The Benchmark. In: AAAI. (2016), 308-314
 - Yu, L., Park, E., Berg, A.C., Berg, T.L.: Visual madlibs: Fill in the blank description generation and question answering. In: Proceedings of the IEEE international conference on computer vision. (2015) 2461-2469
 - Iyyer, M., Manjunatha, V., Guha, A., Vyas, Y., Boyd-Graber, J.L., Daume III, H., Davis, L.S.: The Amazing Mysteries of the Gutter: Drawing Inferences Between Panels in Comic Book Narratives. In: CVPR. (2017) 6478-6487
- ➔ For a rather extensive overview see Pezzelle et al. SiVL 2018