Multimodal Learning and Reasoning

Desmond Elliott, Douwe Kiela and Angeliki Lazaridou

University of Amsterdam, d.elliott@uva.nl
University of Cambridge, douwe.kiela@cl.cam.ac.uk
University of Trento, angeliki.lazaridou@unitn.it

August, 2016

Schedule and Communication

0900-1030	Introduction Grounded Lexical Semantics Referential Grounding
1030-1055	Coffee Break!
1100-1200	Reasoning and Understanding Beyond Words
1200-1230	Final Words and Open Discussion

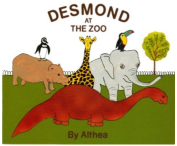
Everyone	#acl2016berlin
Us	@delliott and @aggielaz
Later	`http://multimodalnlp.github.io`

Humans constantly excel in a variety of tasks

Multimodal nature of human intelligence

Humans constantly excel in a variety of tasks

Multimodal nature of human intelligence



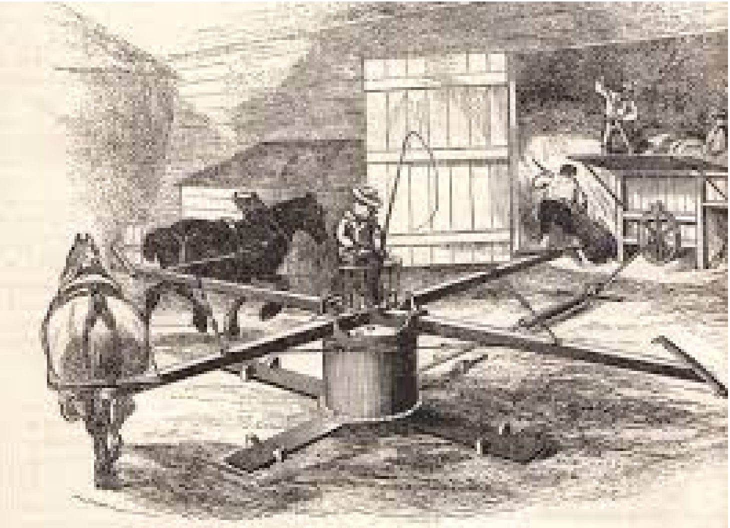
Humans constantly excel in a variety of tasks

Multimodal nature of human intelligence

Humans constantly excel in a variety of tasks

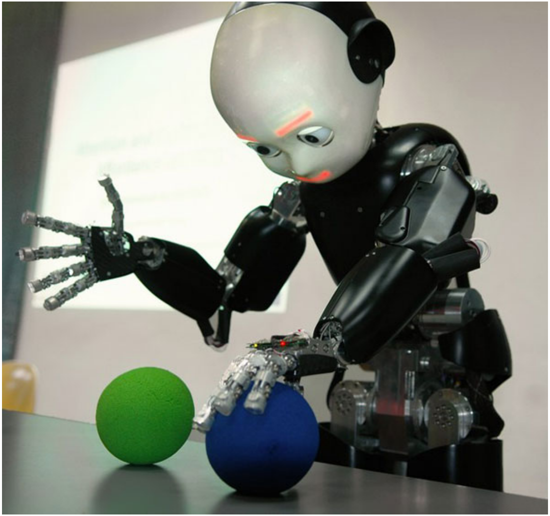
Multimodal nature of human intelligence

Machines are constantly trying to catch up



Machines are constantly trying to catch up

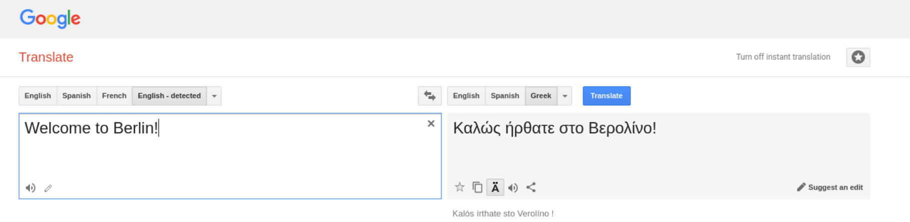
Modalities: vision, haptic, sensors, language



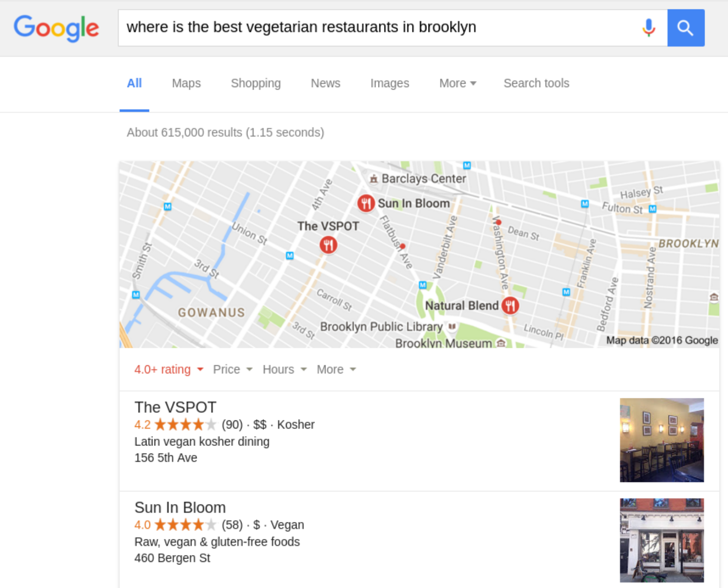
Machines are constantly trying to catch up

Modalities: vision, sensors, GPS

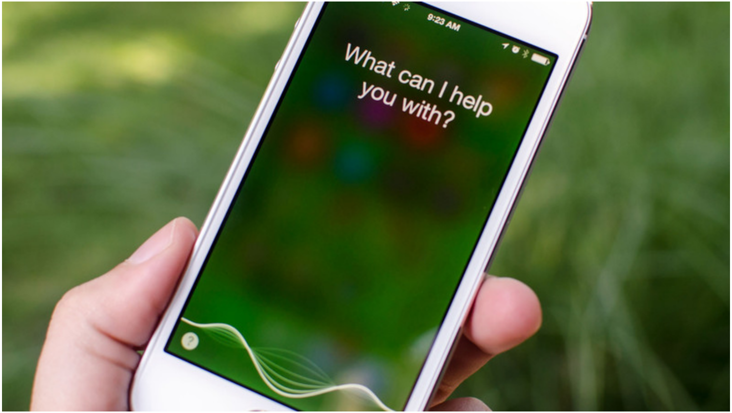
NLP is advancing...



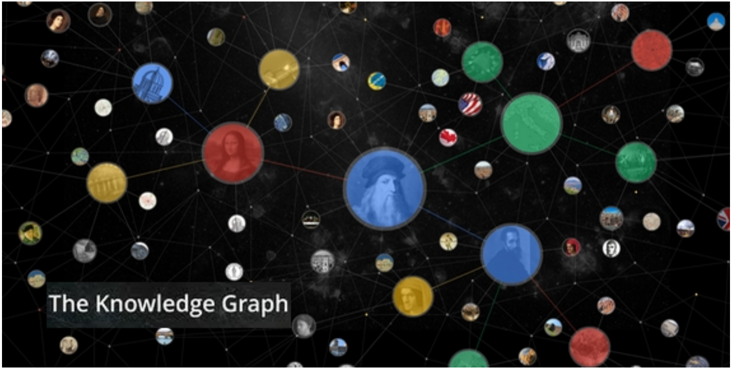
NLP is advancing...



NLP is advancing...



NLP is advancing...

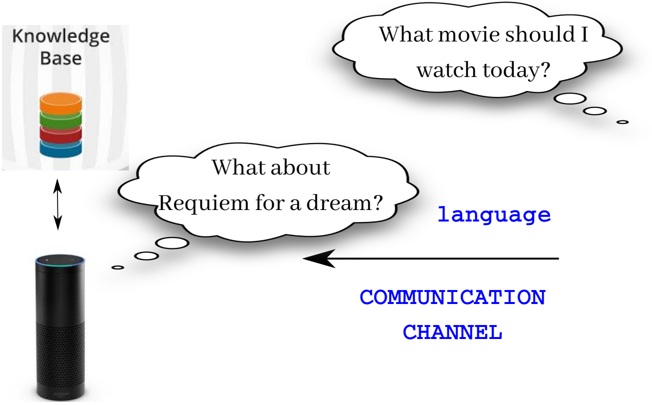


...or maybe not?

Moving beyond the linguistic modality

...or maybe not?

Moving beyond the linguistic modality



...or maybe not?

Moving beyond the linguistic modality

What about Ice Age?

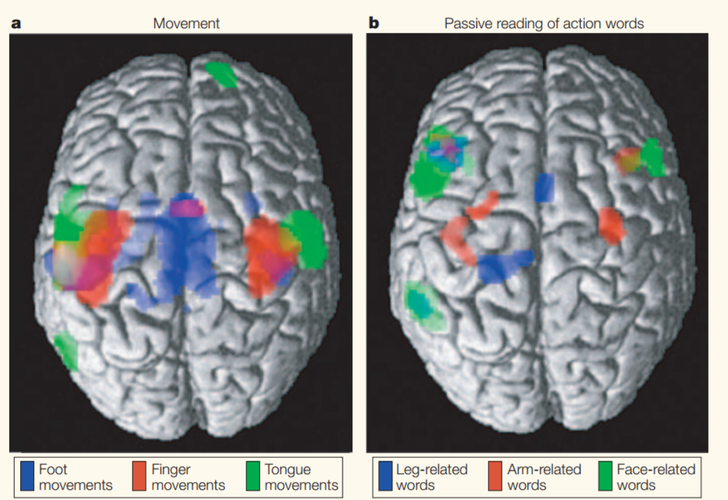
What movie should I watch today?

visual context
language
prosody
facial expressions

COMMUNICATION CHANNEL

Evidence in favor of multimodal language understanding

Motor system activates when reading action words [Pulvermuller, 2005]



Evidence in favor of multimodal language understanding

Purely linguistic or conceptual construction of sentence meaning? [Potter et al., 1986]

Judy needed the  to reach the 

Evidence in favor of multimodal language understanding

Gestures convey information not found in speech [Goldin-Meadow, 2003]

Language can be better understood when presented and interpreted in the context of the world it pertains to.

Multimodality helps with classic NLP tasks

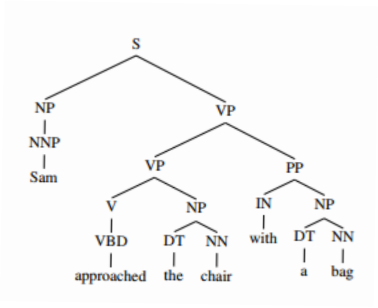
PP attachement disambiguation [Berzak et al., 2015]

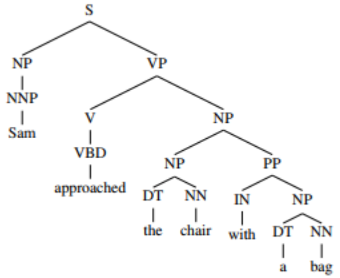
Sam approached the chair with a bag.

Multimodality helps with classic NLP tasks

PP attachement disambiguation [Berzak et al., 2015]

Sam approached the chair with a bag.





Multimodality helps with classic NLP tasks

PP attachement disambiguation [Berzak et al., 2015]

Sam approached the chair with a bag.

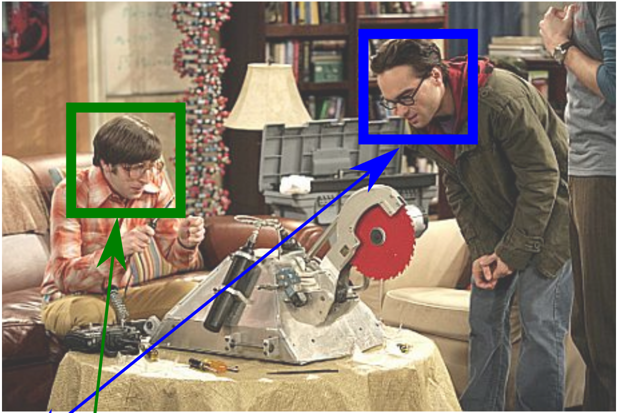
Multimodality helps with classic NLP tasks

Co-reference resolution [Ramanathan et al., 2014]

Leonard looks at the robot, while the only **engineer** in the room fixes it. **He** is amused.

Multimodality helps with classic NLP tasks

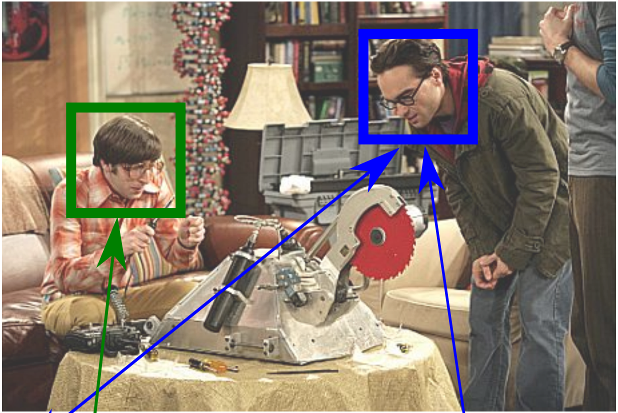
Co-reference resolution [Ramanathan et al., 2014]



Leonard looks at the robot, while the only **engineer** in the room fixes it. **He** is amused.

Multimodality helps with classic NLP tasks

Co-reference resolution [Ramanathan et al., 2014]

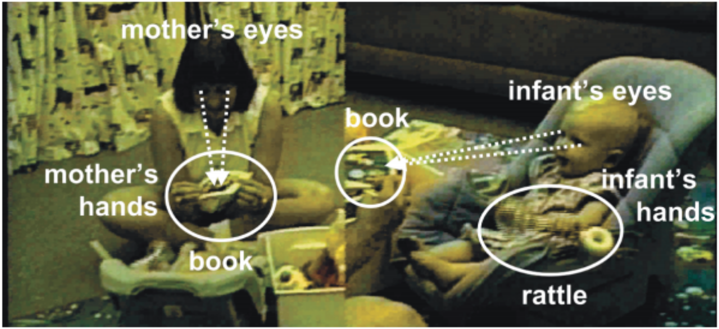


Leonard looks at the robot, while the only **engineer** in the room fixes it. **He** is amused.

Multimodality helps with classic NLP tasks

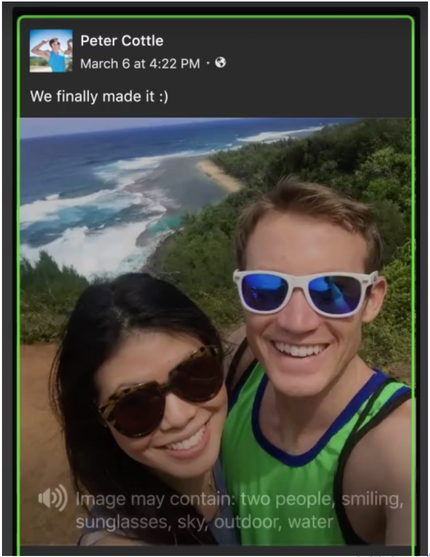
Reference resolution

- [Frank et al., 2013]: social cues (e.g., eye-gaze, body posture)
- [Lazaridou et al., 2016]: social cues + images



When does multimodality make sense?

Assisting visually-impaired people (Facebook)



When does multimodality make sense?

Socially assistive robots that help kids practise their social skills (Robots4Autism)

A Tutorial on Multimodality?

Multimodal NLP is moving beyond an "emerging area" of research:

2011-	V&LNet Vision & Language Workshops
ACL 2013	Visual Features for Lingusitics. Bruni and Baroni.
EACL 2014	Describing Images in Natural Lanugage. Hockenmaier.
CVPR 2015	Vision & Language Workshop
iV&L 2015-16	Vision and Language Summer Schools
NIPS 2015	Multimodal Machine Learning Workshop
MM 2016	Vision and Language Integration Meets Multimedia Fusion
ACL 2016	Multimodal Learning and Reasoning

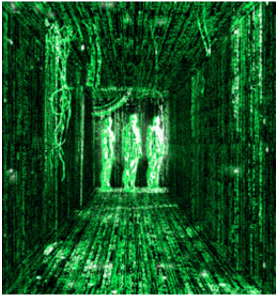
Overview

1. Part 1: Modalities, Representations & Tools
2. Part IIa: Grounded Lexical Semantics
3. Part IIb: Linking words to things
4. Coffee break!
5. Part III: Reasoning and Understanding Beyond Words
6. Final Words

Part 1: Modalities, Representations & Tools

AI's Most Valuable Problem

- Meaning is the "**holy grail**" [Jackendoff, 2002]
- We need to relate semantics to **physical reality** / **sensorimotor experience**.
- **Three levels** of human information processing (Hassabis):
  1. Perceptual input
  2. Conceptual representation
  3. Symbolic reasoning



Most Valuable Problem for AI: *how is it that perceptual input leads to conceptual representations that can be reasoned with?*

Of Tigers and Men

Resources describing **tigers**

Distributional models	live in jungle, can kill, risk extinction

Of Tigers and Men

Resources describing **tigers**

Distributional models	live in jungle, can kill, risk extinction
Perceptual norms	have stripes, have teeth, are orange and black

Resources describing **tigers**

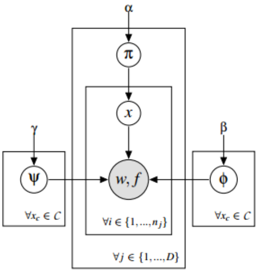
Distributional models	live in jungle, can kill, risk extinction
Perceptual norms	have stripes, have teeth, are orange and black
Perception	

Perceptual Input via Property Norms: Early Examples

[Silberer and Lapata, 2012]

[Andrews et al., 2009]

- Feature-topic model conditions on word-feature pairs from joint corpus



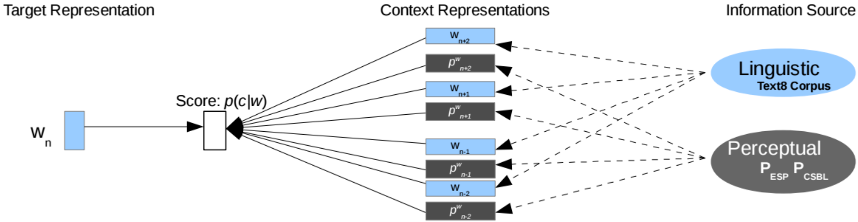
[Johns and Jones, 2012]

- A word's meaning is represented by concatenating its distributional and perceptual representation.
- If no perceptual representation exists, we can infer it, constructing a "global similarity model".

Perceptual Input via Property Norms: Skip-grams

[Hill and Korhonen, 2014]

- Perceptual norms as a proxy for sensorimotor experience **using skip-grams**.



Problems with Perceptual Norms

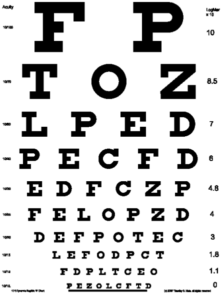
- **Proxy** for real perception
- **Expensive** to obtain
- **Small** datasets (few target cues)
- **Limited** in number (few properties)
- **Mixed**-modality
- People are **bad at listing things**
- **Miss** obvious attributes (e.g. *cats have a neck*)
- Examples of norms:
  - USF norms (association) [Nelson et al., 2004]
  - McRae norms (property) [McRae et al., 2005]
  - CSLB norms (property) [Devereux et al., 2014]

From Perception to Concept Representation

From Perception to Concept Representation

Raw Perceptual Input

- Instead of using norms, use "raw" perceptual input: **images**.
- How do we get **representations**? Two main methods:
    - Bag of visual words [Sivic and Zisserman, 2003]
    - Convolutional neural networks [LeCun et al., 1998]



Bag of visual words

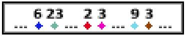
1. Identify keypoints
   1. identify using SIFT [Lowe, 2004]
   2. lay out on dense grid
2. Get local feature descriptors
3. Cluster local descriptors
4. Quantize

Bag of visual words

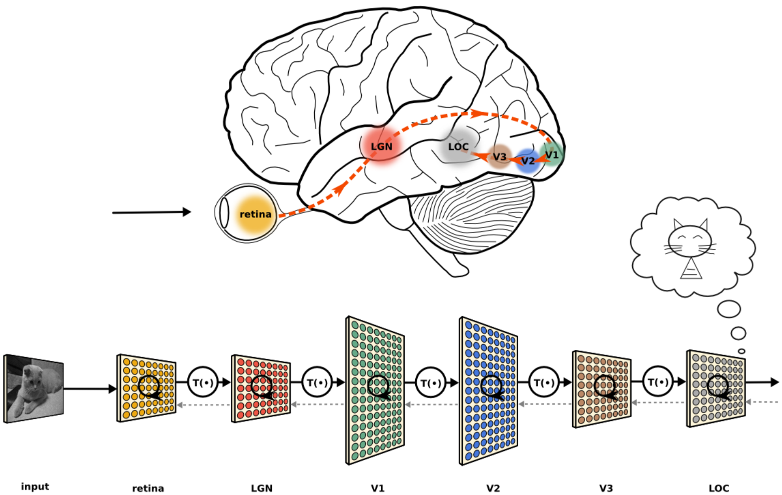
1. Identify keypoints
   1. identify using SIFT [Lowe, 2004]
   2. lay out on dense grid
2. Get local feature descriptors
3. Cluster local descriptors
4. Quantize

Bag of visual words

1. Identify keypoints
   1. identify using SIFT [Lowe, 2004]
   2. lay out on dense grid
2. Get local feature descriptors
3. Cluster local descriptors
4. Quantize



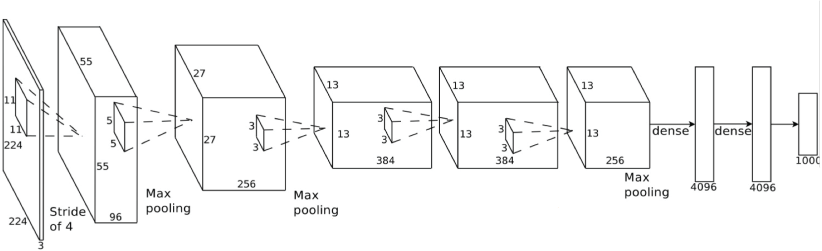
Convolutional Neural Networks: Motivation

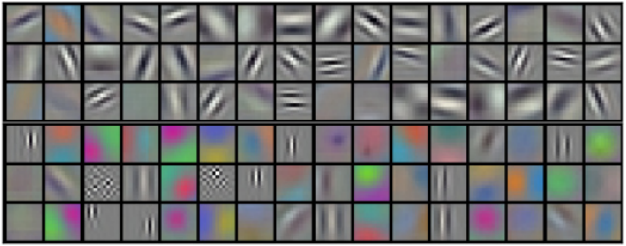


(c) Jonas Kubilias

Convolutional Neural Networks

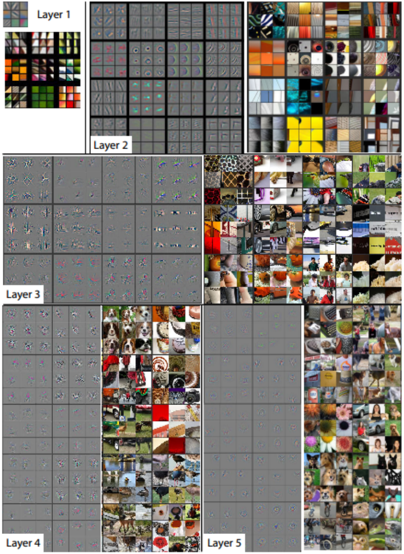
AlexNet [Krizhevsky et al., 2012a]



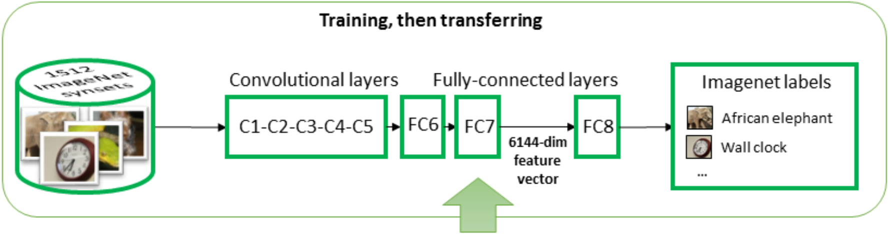


Convolutional Neural Networks: Off-the-shelf

- **State-of-the-art** in most of computer vision [LeCun et al., 2015]
- Off-the-shelf CNN features are **astounding baseline for recognition** [Razavian et al., 2014]
- Work on **visualising and understanding** [Zeiler and Fergus, 2014]



Convolutional Neural Networks: Transferring



1. Train a **convolutional neural network** on a vision task e.g. AlexNet [Krizhevsky et al., 2012b] on ILSVRC [Russakovsky et al., 2015]
2. Do a **forward pass** given an image input
3. **Transfer** one or more layers (e.g. FC_7, or $CONV_5$)

Sources of Image Data

- Different **sources of image data** available
  1. ImageNet
  2. ESP Game Dataset
  3. Wikipedia
  4. News
  5. Image search engines (Google, Bing, Flickr)
  6. MS-COCO
  7. Yahoo 100M
  8. PASCAL VOC
  9. TUHOI
  10. ImageCLEF
  11. ... and many, many more.

Word labels: ImageNet, ESP Game

- Standard datasets of **human-annotated labels**
- ESP: Game with a purpose (GWAP)
- **Advantages**: human-annotated, WordNet-aligned (ImageNet)
- **Disadvantages**: single word labels, low coverage

Joint text and images: Wikipedia, News, Web

- The web contains a plethora of joint image-text data.
- Higher quality: Wikipedia, News
- Lower quality: any web page
- **Advantages**: jointly learnable, easily accessible
- **Disadvantages**: noisy, less descriptive images

Golden Retriever

Origin Scotland

Traits [show]

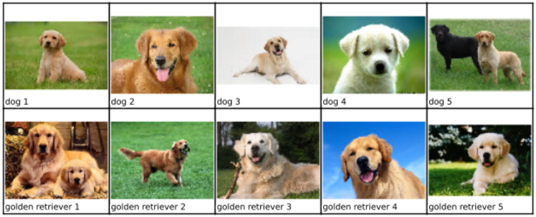
Classification / standards [hide]

FCI	Group 8, Section 1 #111	standard
AKC	Sporting	standard
ANKC	Group 3 (Gun dogs)	standard
CKC	Group 1 – Sporting dogs	standard
KC (UK)	Sporting dog	standard
UKC	Sporting and fishing	

Domestic dog (*Canis lupus familiaris*)

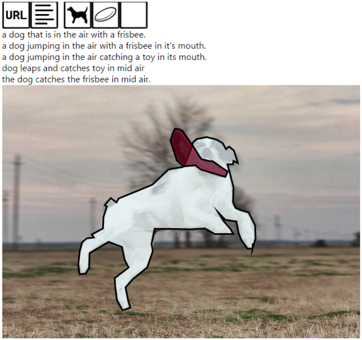
Search engines

- Hybrid: search engines trained on the Web for accurately labelling images
- **Advantages**: massive coverage, easily accessible
- **Disadvantages**: black box



Other labels: Captions, Questions, etc.

- MS-COCO
- Yahoo 100M
- PASCAL VOC
- TUHOI
- ImageCLEF
- **Advantages**: lots of variety, some are huge, annotations are phrases/sentences/paragraphs
- **Disadvantages**: noisy for concept learning, annotator-reliant, often biased



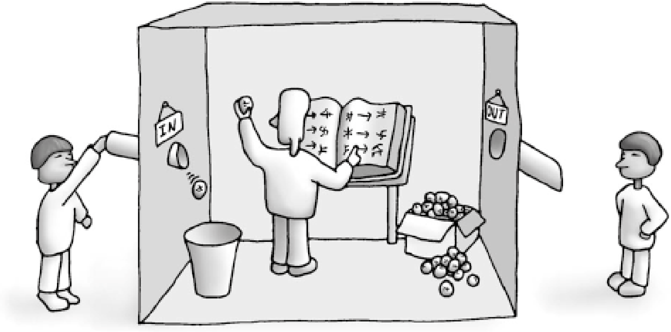
Tools

- Many tools available for extracting representations from images.
- Computer vision:
  - VLFeat (Matlab) - http://www.vlfeat.org
  - OverFeat (C++) - https://github.com/sermanet/OverFeat
  - Caffe (C++/Python) - http://caffe.berkeleyvision.org
  - Cuda-convnet (C++) - https://code.google.com/p/cuda-convnet/
- Multi-modal semantics:
  - MMFeat (Python) - https://github.com/douwekiela/mmfeat
  - VSEM (Matlab) - http://clic.cimec.unitn.it/vsem
- General ML/DL:
  - Torch/Theano/TensorFlow/Keras etc.

Part IIa: Grounded Lexical Semantics

Long history: Symbol grounding problem

[Searle, 1980, Harnad, 1990]



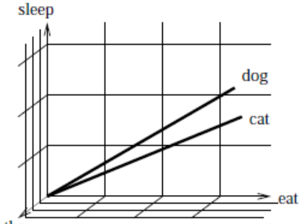
How can you know the meaning of a symbol
if it is defined through other symbols?

Distributional hypothesis

"You shall know a word by the company it keeps"
(Firth, 1957; Harris, 1952)

the [furry **cat** purred] while [the **dog** barked] outside

	purr	bark	fur	animal	landing	space	sleep	eat
dog	2	19	15	30	1	2	10	23
cat	23	5	19	25	0	1	34	19
moon	0	2	0	1	25	17	7	1



(Clark '14)

(Baroni, V&L '15)

Grounding problem in semantics: Meaning is grounded

Glenberg & Robertson 2000; Barsalou 2008; Andrews et al. 2009; Baroni et al. 2010; Riordan & Jones 2011; Bruni et al. 2014

democracy

/dɪˈmɒkrəsi/

noun

a system of government by the whole population or all the eligible members of a state, typically through elected representatives.
"a system of parliamentary democracy"

cat[1]

/kat/

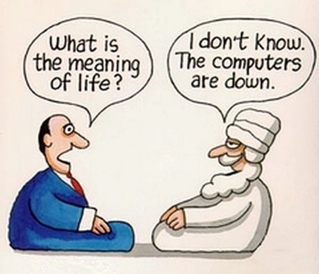
noun

1. a small domesticated carnivorous mammal with soft fur, a short snout, and retractile claws. It is widely kept as a pet or for catching mice, and many breeds have been developed.

Meaning is grounded in sensori-motor experience!

Grounding problem in semantics: Grounding helps

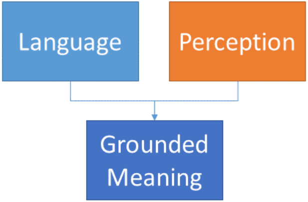
- Grounding helps for:
  - Similarity and relatedness
  - Concept categorization
  - Compositionality
  - Bilingual lexicon induction
  - Lexical entailment
  - Metaphor detection
  - Visual information retrieval



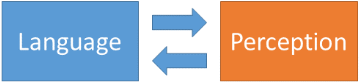
Grounding (we believe) leads to more "human" meaning representations

Grounding at different levels of meaning

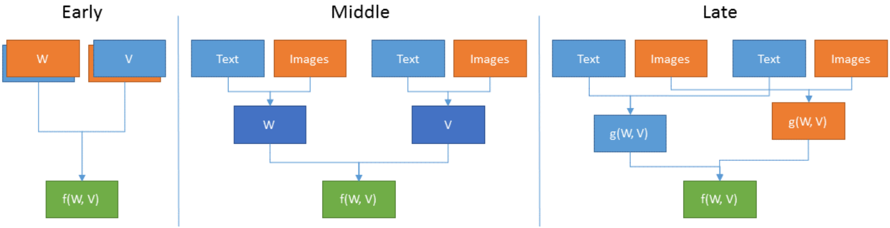
- **Representational** grounding
  - Multi-modal semantics: Representing the grounded meaning of a word
  - Frege's *Sinn* (sense)
  - **Core issue**: fusion



- **Referential** grounding
  - Cross-modal semantics: Determining the referent that a word denotes
  - Frege's *Bedeutung* (reference)
  - **Core issue**: mapping



Multi-modal fusion



- We need to perform **fusion** of textual and perceptual information.
  - Early: learn jointly, then compute function
  - Middle: learn separately, then combine, then compute function
  - Late: learn separately, compute function individually and combine function outputs

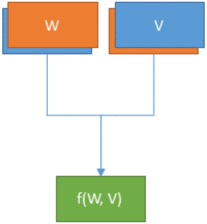
Evaluating grounded representations

automobile	car	1.00	$sim(\vec{v}_{automobile}, \vec{v}_{car})$
eagle	feathers	0.88	$sim(\vec{v}_{eagle}, \vec{v}_{feathers})$
...	...	...	...
bakery	zebra	0.00	$sim(\vec{v}_{bakery}, \vec{v}_{zebra})$

- **Similarity and relatedness** (Spearman correlation)
  - MEN
  - SimLex-999
  - WordSim353
  - ... many more ...
- Great results with multi-modal semantics

Early fusion: Topic models

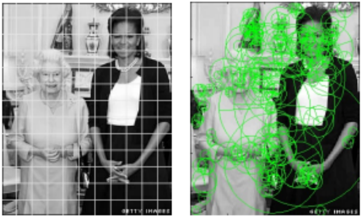
[Feng and Lapata, 2010b, Roller and Schulte im Walde, 2013]



- Topic model of multi-modal documents using bag of visual words (SIFT/SURF)
- May also include perceptual norms

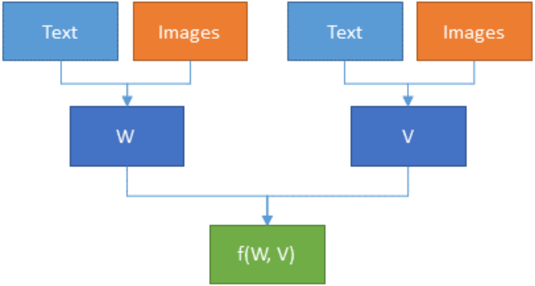
Michelle Obama fever hits the UK

In the UK on her first visit as first lady, Michelle Obama seems to be making just as big an impact. She has attracted as much interest and column inches as her husband on this London trip; creating a buzz with her dazzling outfits, her own schedule of events and her own fanbase. Outside Buckingham Palace, as crowds gathered in anticipation of the Obamas' arrival, Mrs Obama's star appeal was apparent.

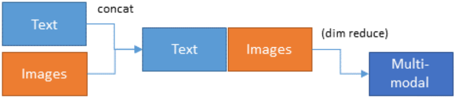


Mid fusion: Early work

[Bruni et al., 2011, Leong and Mihalcea, 2011b, Bruni et al., 2012, Bruni et al., 2014]



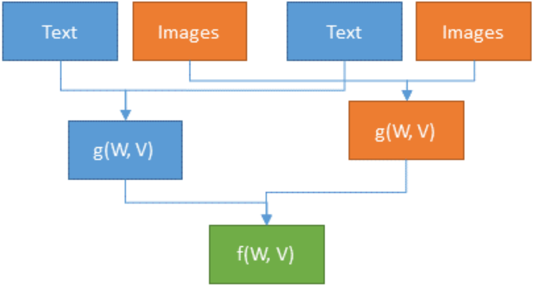
1. Combine uni-modal representations



2. Compute function over multi-modal inputs, e.g. cosine

Late fusion: Early work

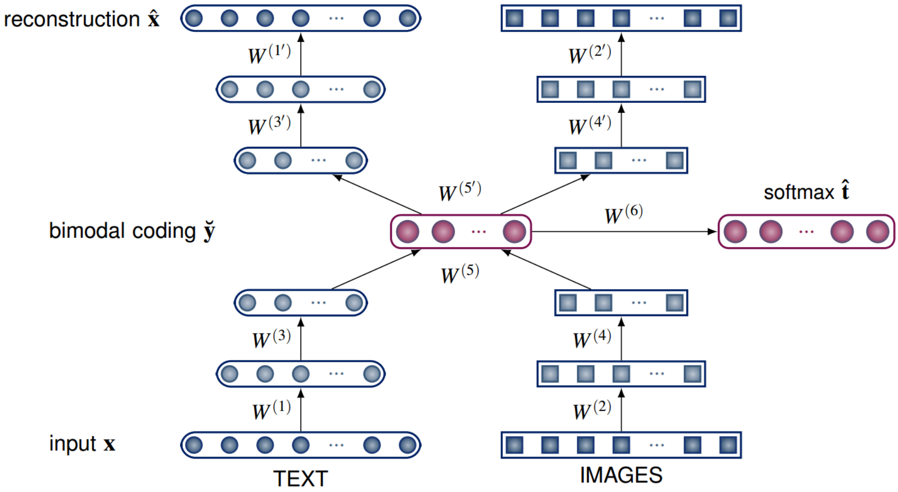
[Leong and Mihalcea, 2011a]



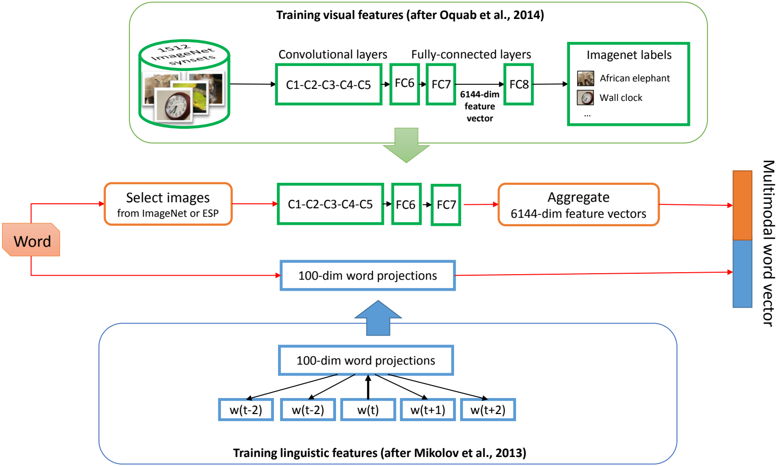
1. Compute uni-modal function over the inputs, e.g. cosine
2. Combine the function outputs using another function

Grounded meaning with autoencoders

[Silberer and Lapata, 2014]

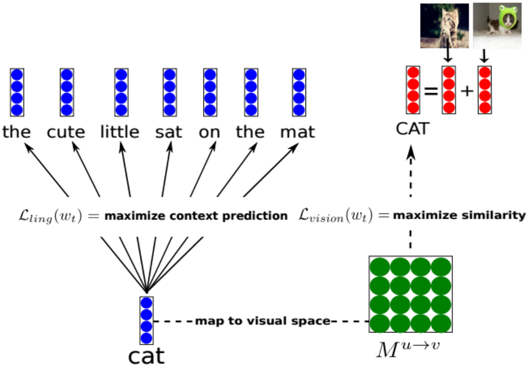


Improved multi-modal semantics with image embeddings
[Kiela and Bottou, 2014]



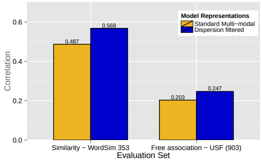
Multi-modal skip-gram

[Lazaridou et al., 2015c]



$$J = \frac{1}{T} \sum_{t=1}^{T} \left(\mathcal{L}_{ling}(w_t) + \mathcal{L}_{vision}(w_t) \right)$$

- Predict concreteness/abstractness of concepts based on images
- Compare elephant and happiness
- **Image dispersion**-based filtering



$$d(w) = \frac{2}{n(n-1)} \sum_{i<j\leq n} 1 - cos(\vec{w}_i, \vec{w}_j)$$

Applications: Selectional preferences
[Bergsma and Goebel, 2011]

- Use visual properties for predicting selectional preference
- In their DSP model, introduce textual as well as visual features.
- Get images from Flickr and Google
- **Multi-modal works best**

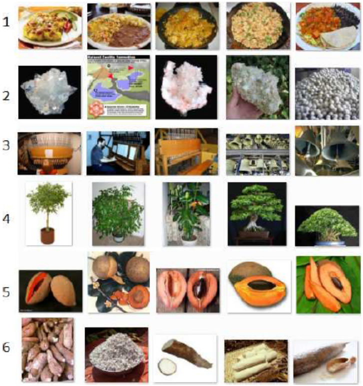
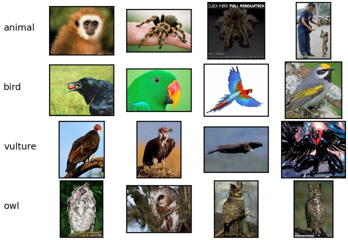


Figure 1: Which out-of-vocabulary nouns are plausible direct objects for the verb *eat*? Each row corresponds to a noun: 1. *migas*, 2. *zeolite*, 3. *carillon*, 4. *ficus*, 5. *mamey* and 6. *manioc*.

Applications: Visual lexical entailment
[Kiela et al., 2015c]

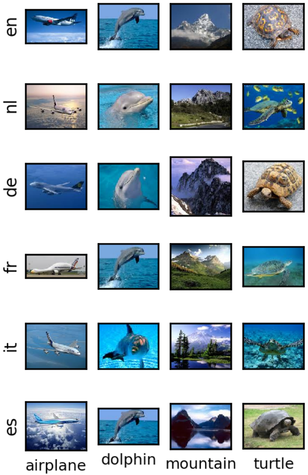
- Lexical entailment:
  Animal ⇒ Bird ⇒ Raptor ⇒ Vulture
- Idea: exploit **generality** of images from Google Images
- **Multi-modal works best**



Applications: Visual bilingual lexicon induction

[Bergsma and Van Durme, 2011, Kiela et al., 2015d, Vulić et al., 2016]

- Bilingual lexicon induction:
  Airplane ⇔ Avion ⇔ Flugzeug ⇔ Vliegtuig
- Idea: exploit cross-lingual **similarity** of images from Google Images
- **Multi-modal works best**



Applications: Metaphor detection

[Shutova et al., 2016]

[Mohammad et al., 2016] - SV/VO	
blister foot	literal
blister administration	metaphorical
blur vision	literal
blur distinction	metaphorical
[Tsvetkov et al., 2014] - AN	
cold beer	literal
cold heart	metaphorical
foggy morning	literal
foggy brain	metaphorical

- Task: classify S-V, V-O and A-N pairs according to metaphoricity
- **Multi-modal works best**

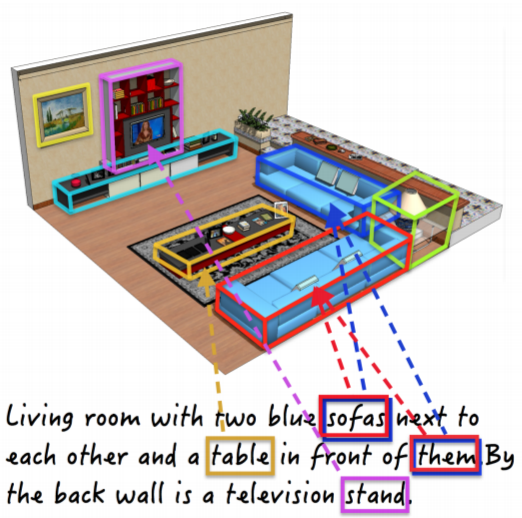
Next: Important questions

- **Why** do multi-modal representations work so well?
- Is it just **extra information**, is it **complementary**, is it **fundamentally different**?
- How about **other modalities**? And **other tasks**?
- Can we do multi-modal **composition**? What does that even mean?

Only the beginning of this field: **many exciting things left to do!**

Part IIb: Linking words to things

Referential Grounding: Linking words and real world



Lack of *reference* in semantics

Natural language is, fundamentally, a means to **communicate**. Our words must be able to **refer** to the objects, properties and events in the outside world.

Lack of *reference* in semantics

Natural language is, fundamentally, a means to **communicate**. Our words must be able to **refer** to the objects, properties and events in the outside world.

- Current models of meaning are purely **language-internal**.
- NLP agents cannot reason about simple statements regarding **the real world** (*"Is there a cat in the room?"*)

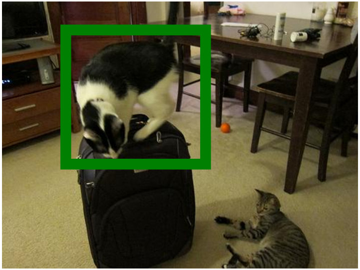
Why should we care about reference?

Interpreting linguistic expressions requires more than just identifying **linguistic relations** between words.

Baroni, 2016, p4

Crucial for **Referring Expression Generation**[1]

[Dale and Reiter, 1995, Mitchell et al., 2010, Kazemzadeh et al., 2014]



→ My cat is the one on top of the luggage.

[1] A comprehensive study at [Krahmer and Van Deemter, 2012]

Why should we care about reference?

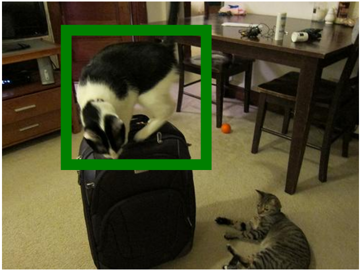
Interpreting linguistic expressions requires more than just identifying **linguistic relations** between words.

Baroni, 2016, p4

Crucial for **Reference Resolution**

[Roy, 2002, Matuszek et al., 2012, Schlangen et al., 2015]

My cat is the one
on top of the luggage.

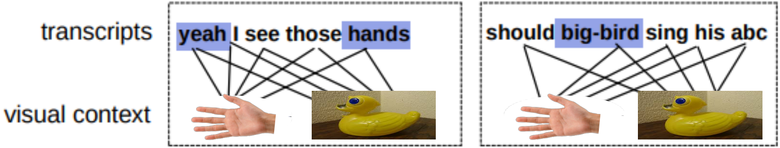


Why should we care about reference?

Interpreting linguistic expressions requires more than just identifying **linguistic relations** between words.

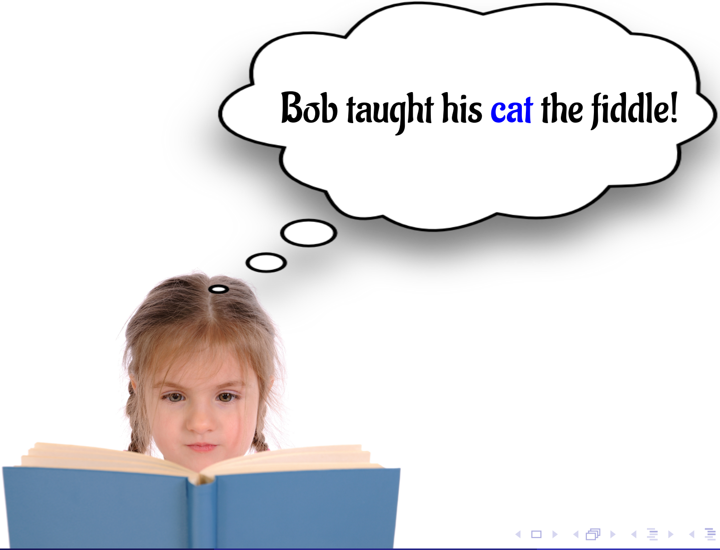
Baroni, 2016, p4

Crucial for **Cross-situational Language Learning** [Siskind, 1996, Yu and Ballard, 2004, Fazly et al., 2010, Chrupała et al., 2015, Lazaridou et al., 2016]



Humans performing referential grounding

"Visualizing" the meaning of familiar concepts

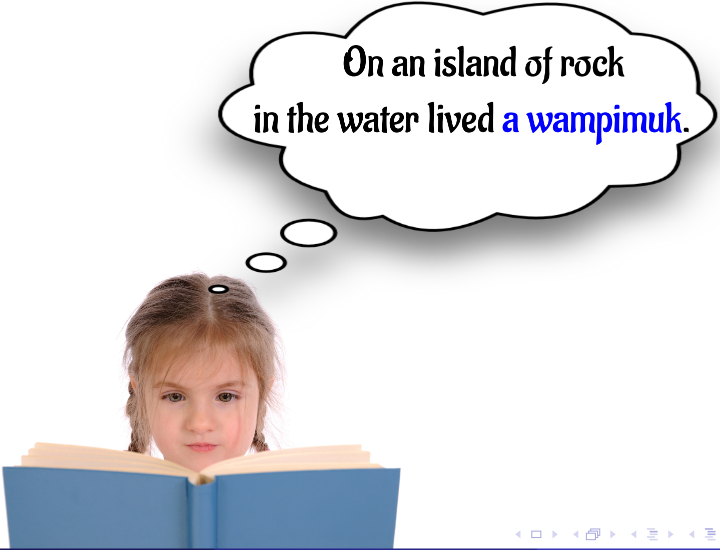


Humans performing referential grounding

"Visualizing" the meaning of familiar concepts

Humans performing referential grounding

Draw inferences for novel concepts



Humans performing referential grounding

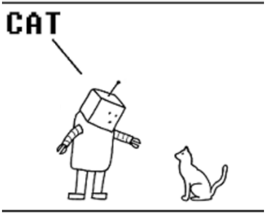
Draw inferences for novel concepts

Humans performing referential grounding

Draw inferences for novel concepts

Machines performing referential grounding

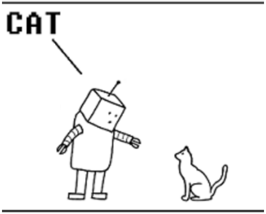
"cats" vs "wampimuks"

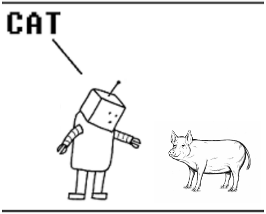


- For familiar concepts (e.g., *cat*), build a **naive** pipeline based on ConvNets
  - **Pros**: High accuracy for familiar concepts (pre-trained ConvNets predicts 1000 concepts)
  - **Cons**: "Limited" labeled datasets,

Machines performing referential grounding

"cats" vs "wampimuks"

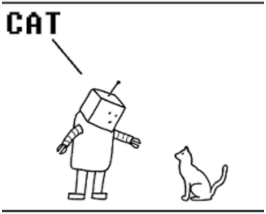


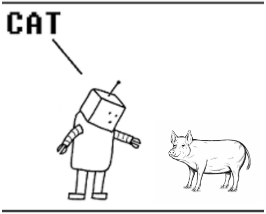


- For familiar concepts (e.g., *cat*), build a **naive** pipeline based on ConvNets
  - **Pros**: High accuracy for familiar concepts (pre-trained ConvNets predicts 1000 concepts)
  - **Cons**: "Limited" labeled datasets, no generalization to new concepts

Machines performing referential grounding

"cats" vs "wampimuks"





- For familiar concepts (e.g., *cat*), build a **naive** pipeline based on ConvNets
  - **Pros**: High accuracy for familiar concepts (pre-trained ConvNets predicts 1000 concepts)
  - **Cons**: "Limited" labeled datasets, no generalization to new concepts

We need a general mechanism able to handle both familiar and novel concepts

Humans bridge the gap between linguistic and visual experiences

visual experience

Humans bridge the gap between linguistic and visual experiences

visual experience about **cats**

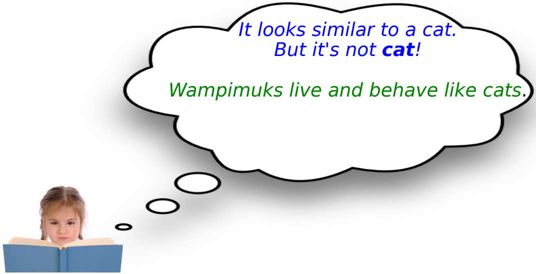
visual experience

Humans bridge the gap between linguistic and visual experiences

visual experience about **cats**

visual experience

linguistic knowledge about **wampimuks and cats**



Humans bridge the gap between linguistic and visual experiences

visual experience about **cats**

visual experience

linguistic knowledge about

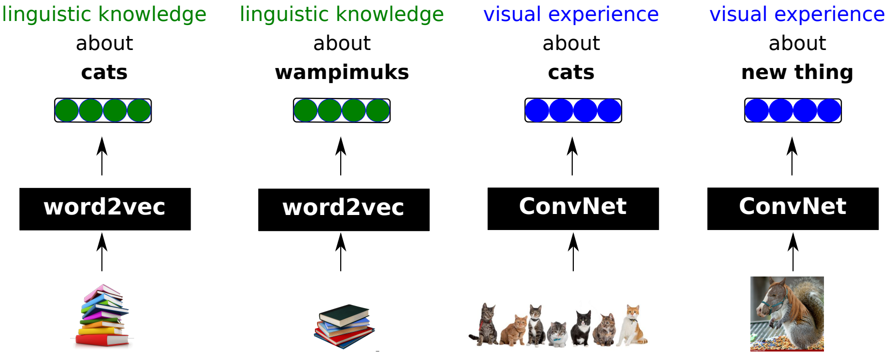
wampimuks and cats

It looks similar to a cat.
But it's not ***cat****!*

Wampimuks live and behave like cats.

It should be a ***wampimuk****!*

Machines representing concepts in a vector space



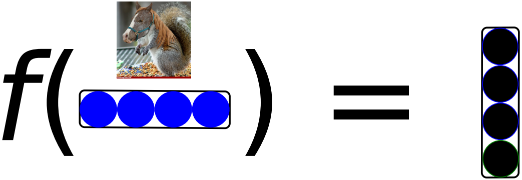
Cross-modal mapping

The heart of the problem [...] is one of **translation**: in order to talk about what we see, information provided by the visual system must be translated into a form compatible with the information used by the language system."

Jackendoff, 1987, p90

Cross-modal mapping

Definition

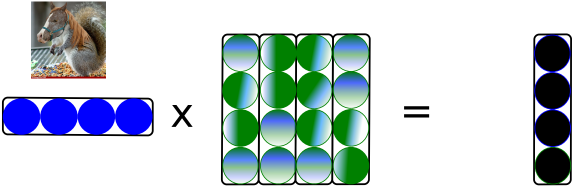


visual vector

predicted linguistic vector

Cross-modal mapping

Example: Linear Mapping

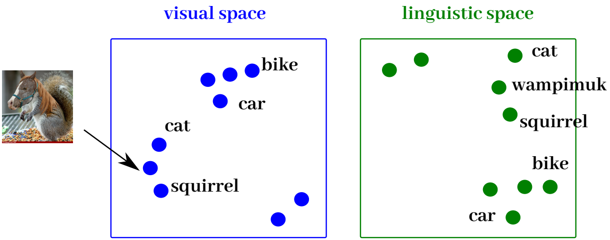


visual vector

cross-modal mapping

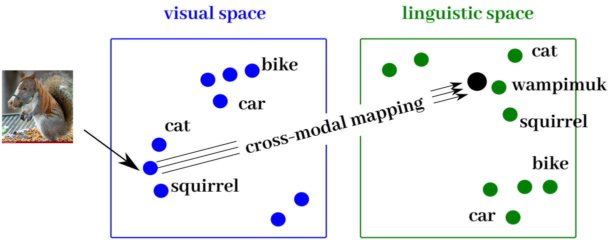
predicted linguistic vector

Referential grounding in vector space through cross-modal mapping



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

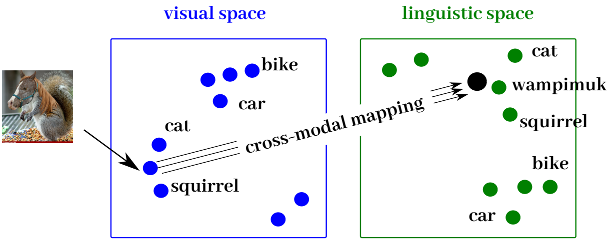
Referential grounding in vector space through cross-modal mapping



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

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

Referential grounding in vector space through cross-modal mapping

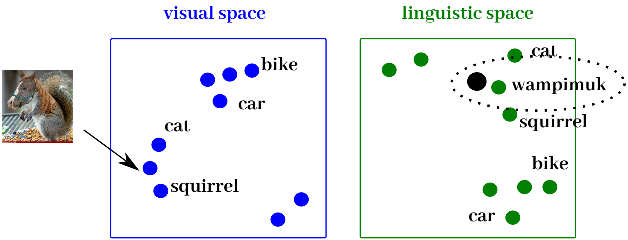


Step 1 Obtain "**parallel data**" of linguistic and visual vectors of concepts.

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

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

Referential grounding in vector space through cross-modal mapping



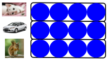
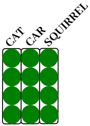
Step 1 Obtain "**parallel data**" of linguistic and visual vectors of concepts.

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**

Cross-modal mapping
Training

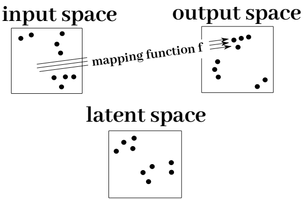
$$\operatorname*{argmin}_{\theta} \operatorname{loss}\Big(f_\theta(\ $$  $$\ ),\ $$  $$\Big)$$

training input space

training output space

- **f**: function paremetrized by weights θ that transforms a visual to a linguistic vector (e.g., linear map)
- **loss**: e.g., *L2* distance, *cosine distance*

Variations for cross-modal mapping[1]



	mapping function	loss	output space
[Socher et al., 2013]	2-layer NN	L2	linguistic
[Frome et al., 2013]	linear map	ranking	linguistic
[Norouzi et al., 2014]	-	-	linguistic
[Lazaridou et al., 2014]	CCA	-	linguistic visual
[Weston et al., 2011]	linear map	ranking	latent
[Srivastava and Salakhutdinov, 2012]	deep boltzmann machines		latent

[1] Recent review by [Wang et al., 2016]

Tasks: Zero-shot Object recognition

Recognizing new concepts by leveraging semantic/linguistic regularities with known concepts.

	Frome et al., 2014	**Softmax over 1k labels**
	eyepiece, ocular Polaroid compound lens **telephoto lens, zoom lens** rangefinder, range finder	typewriter keyboard tape player reflex camera CD player space bar
	fruit pineapple **pineapple plant, Ananas .** sweet orange sweet orange tree, ...	pineapple, ananas coral fungus artichoke, globe artichoke sea anemone, anemone cardoon

Tasks: Zero-shot Object recognition
[Frome et al., 2013], [Socher et al., 2013]

Recognizing new concepts by leveraging semantic/linguistic regularities with known concepts.

	Frome et al., 2014	Softmax over 1k labels
	eyepiece, ocular Polaroid compound lens **telephoto lens, zoom lens** rangefinder, range finder	typewriter keyboard tape player reflex camera CD player space bar
	fruit pineapple **pineapple plant, Ananas .** sweet orange sweet orange tree, ...	pineapple, ananas coral fungus artichoke, globe artichoke sea anemone, anemone cardoon

- 100 labels: 32% Precision@1
- 21k labels: 1% Precision@1

Nearest neighbors analysis

Results on ESP-game dataset

target concept	predicted concept in embedding space	
jellyfish	anemone, jellyfish, seashell	*co-hyponymy*
cow	bison, elephant, baboon	*co-hyponymy*
phone	headset, smartphone, microphone	*meronymy*
instrument	sitar, percussion, accordion	*hyponymy*

How to improve performance of cross-modal mapping (1)

- Inherent properties of **output space** affecting performance
  - traditional word embeddings better *relatedness* vs *similarity* [Kiela et al., 2015b]

Concept	Nearest Neighbors
cat	cats, dogs, scaredy, feline
bike	bikes, bicycle, motorcycle, motorbike

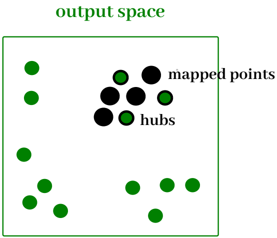
Nearest neighbor queries from the best *predict* CBOW space of [Baroni et al., 2014]

How to improve performance of cross-modal mapping (1)

- Inherent properties of **output space** affecting performance
  - traditional word embeddings better *relatedness* vs *similarity* [Kiela et al., 2015b]
  - 1-vector-per-token resulting in ambiguities

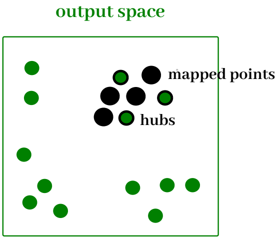
Concept	Nearest Neighbors
cat	cats, dogs, scaredy, feline
bike	bikes, bicycle, motorcycle, motorbike
chair	vice-chair, vice-chairs, co-chair, vice-chairman

Nearest neighbor queries from the best *predict* CBOW space of [Baroni et al., 2014]



- Problem: “Hubs” attract near them predicted points [Radovanović et al., 2010]
  - examples of hubs: smilodon, pintle, handwheel
  - L2 loss for mapping particularly affected by hubness [Shigeto et al., 2015]

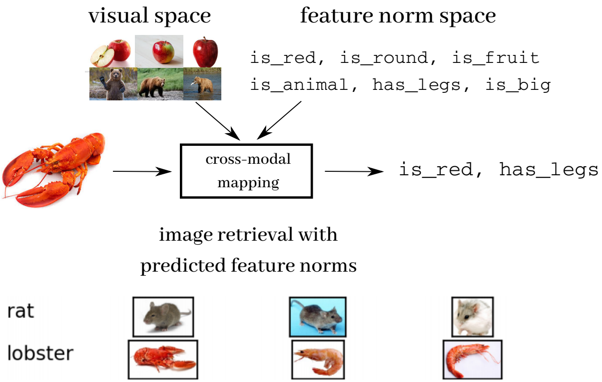
How to improve performance of cross-modal mapping (2)



- Problem: “Hubs” attract near them predicted points [Radovanović et al., 2010]
  - examples of hubs: smilodon, pintle, handwheel
  - L2 loss for mapping particularly affected by hubness [Shigeto et al., 2015]
- Solutions:
  - [Dinu et al., 2015]: use of **globally corrected** nearest neighbor retrieval – downplaying importance of hubs
  - [Lazaridou et al., 2015a]: use of **ranking** instead of L2 loss

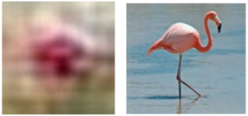
Tasks: Expanding feature norms
[Bulat et al., 2016]

Automatically enlarging coverage of feature norms by mapping visual vectors of novel entries.

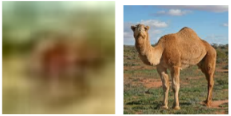


Tasks: Computational Imagery from word embeddings
[Lazaridou et al., 2015b]

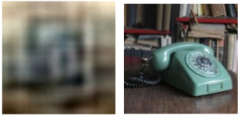
Mapping the word vector of an “unseen” concept onto the *visual* space and then onto the *pixel* space.



flamingo



camel

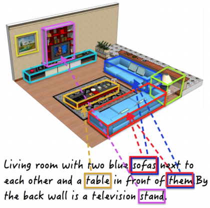


telephone

ambulance

(Not so) Final words

- Referential grounding in vector space through cross-modal mapping
    - A general way to link words to things in the real world
- Moving away from stand-alone architectures to build-in components



Coffee break!

Part III: Reasoning and Understanding Beyond Words

The Need for Reasoning and Understanding

Humans experience the world in a physically embedded setting

The Need for Reasoning and Understanding

Humans experience the world in a physically embedded setting

The Need for Reasoning and Understanding

Not representative example of an actual baby

Humans experience the world in a physically embedded setting

Credit: Stella Frank

Two Tasks for Reasoning and Understanding

1. Image Description

→ A man is pulling off a trick on a snowboard

Two Tasks for Reasoning and Understanding

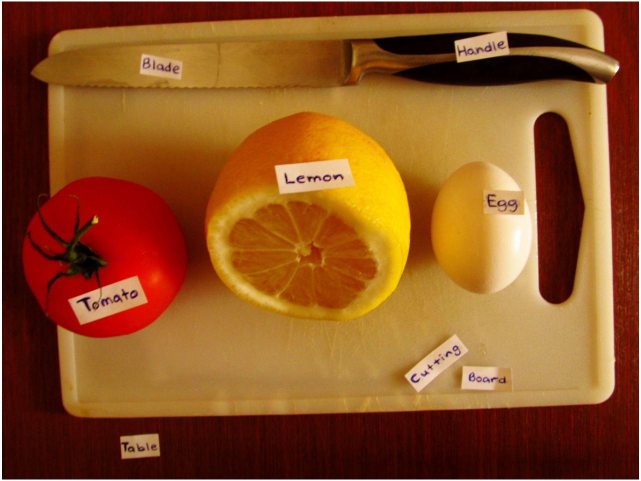
1. Image Description

→ A man is pulling off a trick on a snowboard

2. Visual Question Answering

→ Yellow

What colour is the moustache made of?

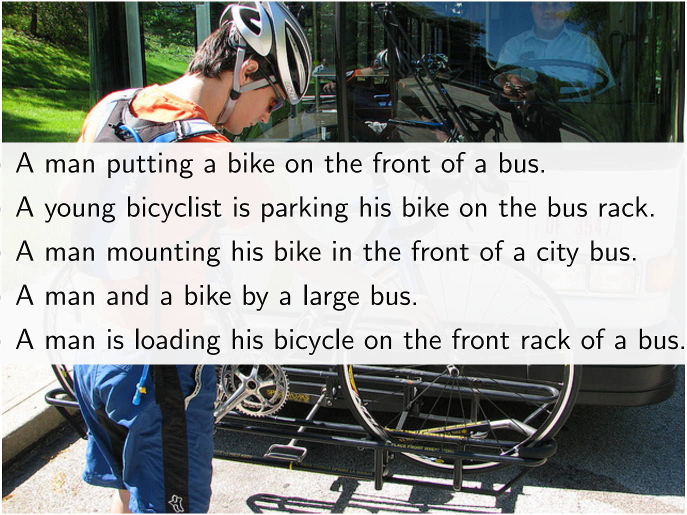


https://www.flickr.com/photos/59152532@N05/14260478426

How would you describe this image?

http://mscoco.org/explore/?id=256981

How would you describe this image?



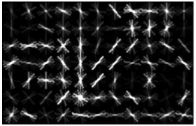
- A man putting a bike on the front of a bus.
- A young bicyclist is parking his bike on the bus rack.
- A man mounting his bike in the front of a city bus.
- A man and a bike by a large bus.
- A man is loading his bicycle on the front rack of a bus.

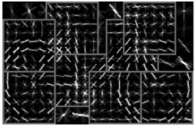
http://mscoco.org/explore/?id=256981

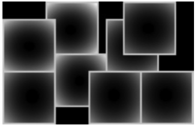
Datasets

	Images	Descriptions	Judgements	Objects
Pascal1K [Rashtchian et al., 2010]	1,000	5	No	No
VLT2K [Elliott and Keller, 2013]	2,424	3	Partial	Partial
Flickr8K [Hodosh et al., 2013]	8,108	5	Yes	No
AbstractScenes [Zitnick and Parikh, 2013]	10,000	6	No	Yes
IAPR-TC12 [Grubinger et al., 2006]	20,000	1–5 En & De	No	Yes
Flickr30K [Young et al., 2014]	31,783	5	No	Partial
Multi30K [Elliott et al., 2016]	31,783	5 En & 6 De	No	Yes
MSCOCO [Chen et al., 2015]	164,062	5	No	Partial

- Early approaches are unified by:
  - SIFT feature vectors [Lowe, 2004]
  - Deformable Parts Object Detections [Felzenszwalb et al., 2008]





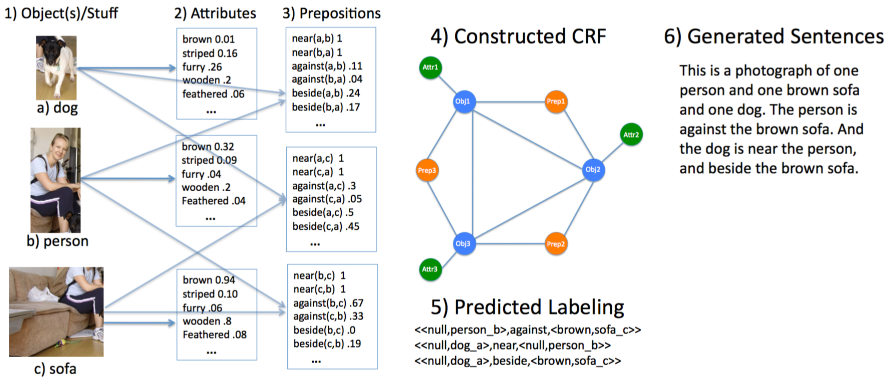


  - Template-based language generation

IMG → DT SUBJ VB OBJ

A person is riding a bike

Objects, Attributes and Prepositions
[Kulkarni et al., 2011]



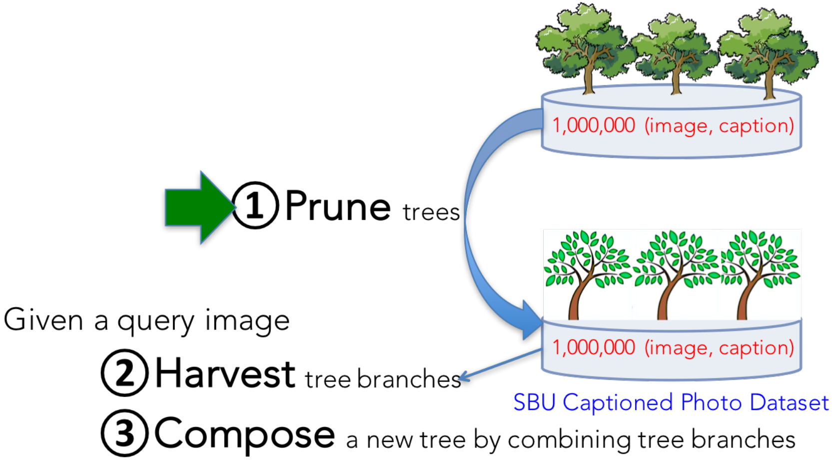
Are we making progress?

[Kulkarni et al., 2011]

There are two aeroplanes.
The first shiny aeroplane is near
the second shiny aeroplane.

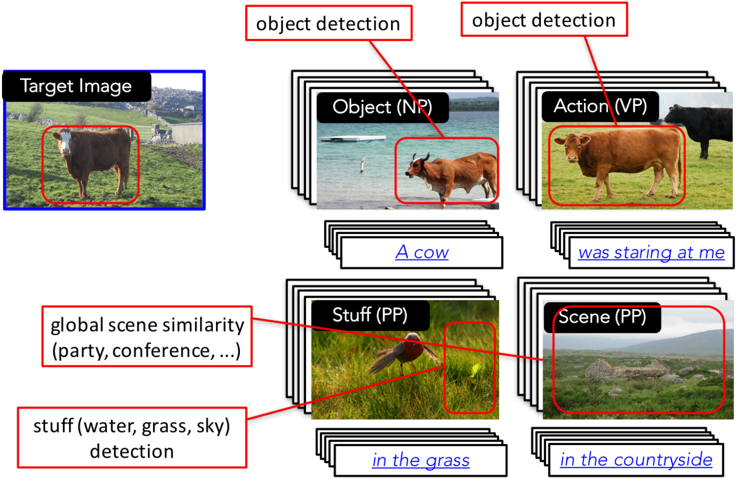
Early Approaches - TreeTalk

[Kuznetsova et al., 2012, Kuznetsova et al., 2014]



Early Approaches - TreeTalk

[Kuznetsova et al., 2012, Kuznetsova et al., 2014]



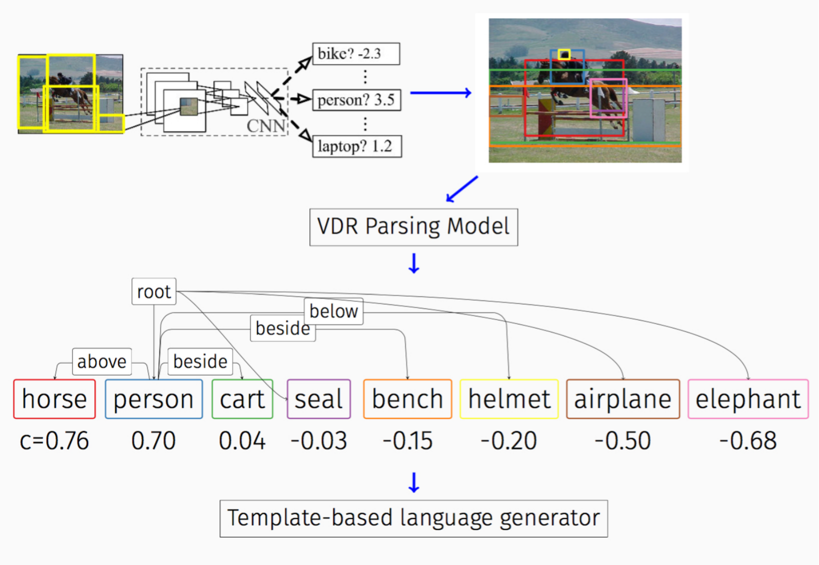
Are we making progress?
[Kuznetsova et al., 2012]

This is a photo of this bird hopping around eating things off of the ground by river.

2011 - - - → 2012

Spatial Relations and Verb Predictions

[Elliott and Keller, 2013, Elliott and de Vries, 2015]



Are we making progress?
[Elliott and Keller, 2013]

A man is holding a phone. A wall is beside a sign.

2011 - - - → 2012 - - - → 2013

An Overview of Early Approaches

Planning & realisation

[Feng and Lapata, 2010a]
[Mitchell et al., 2012]
[Kuznetsova et al., 2012]
[Kuznetsova et al., 2014]

Space and/or Attributes

[Farhadi et al., 2010]
[Kulkarni et al., 2011]
[Elliott and Keller, 2013]
[Yatskar et al., 2014]

Abstract Scenes

[Zitnick and Parikh, 2013]
[Ortiz et al., 2015]

Transfer-based

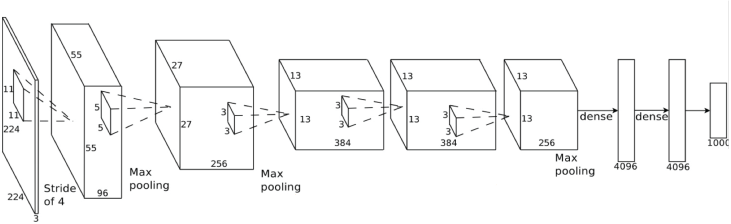
[Ordonez et al., 2011]
[Mason and Charniak, 2014]

External linguistic resources

[Li et al., 2011]
[Yang et al., 2011]

Recent Approaches

- Unified by advances in convolutional neural networks [Krizhevsky et al., 2012a, Simonyan and Zisserman, 2015, He et al., 2015]

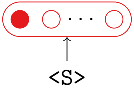


and Recurrent Neural Network language modelling

- New focus on architecture engineering

Convolutional Neural Network - Recurrent Neural Network

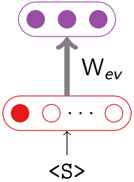
[Vinyals et al., 2015, Karpathy and Fei-Fei, 2015]



Input

Convolutional Neural Network - Recurrent Neural Network

[Vinyals et al., 2015, Karpathy and Fei-Fei, 2015]



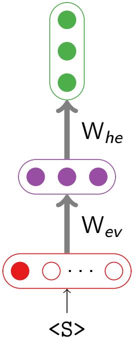
Embeddings

$e_i = x_i \cdot W_{ev}$

Input

Convolutional Neural Network - Recurrent Neural Network

[Vinyals et al., 2015, Karpathy and Fei-Fei, 2015]



Recurrent

$h_i = f(h_{i-1}, e_i)$

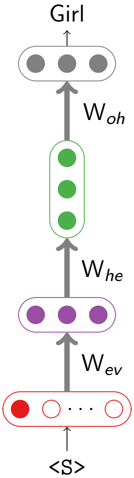
Embeddings

$e_i = x_i \cdot W_{ev}$

Input

Convolutional Neural Network - Recurrent Neural Network

[Vinyals et al., 2015, Karpathy and Fei-Fei, 2015]



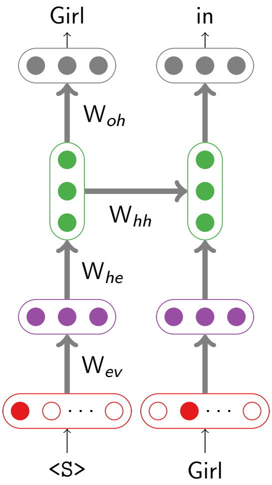
Prediction
softmax(o_i)

Output
$o_i = h_i \cdot W_{oh}$

Recurrent
$h_i = f(h_{i-1}, e_i)$

Embeddings
$e_i = x_i \cdot W_{ev}$

Input



Prediction
softmax(o_i)

Output
$o_i = h_i \cdot W_{oh}$

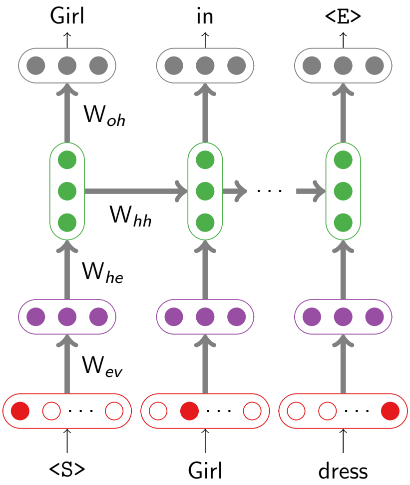
Recurrent
$h_i = f(h_{i-1}, e_i)$

Embeddings
$e_i = x_i \cdot W_{ev}$

Input

Convolutional Neural Network - Recurrent Neural Network

[Vinyals et al., 2015, Karpathy and Fei-Fei, 2015]



Prediction
softmax(o_i)

Output
$o_i = h_i \cdot W_{oh}$

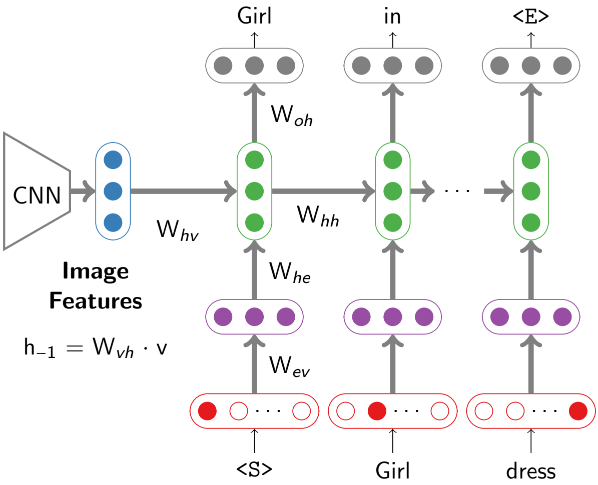
Recurrent
$h_i = f(h_{i-1}, e_i)$

Embeddings
$e_i = x_i \cdot W_{ev}$

Input

Convolutional Neural Network - Recurrent Neural Network

[Vinyals et al., 2015, Karpathy and Fei-Fei, 2015]



Prediction
$\mathrm{softmax}(o_i)$

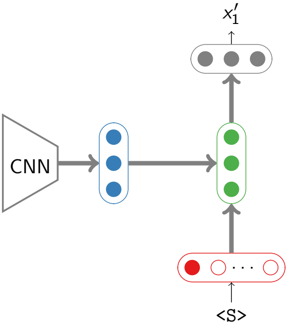
Output
$o_i = h_i \cdot W_{oh}$

Recurrent
$h_i = f(h_{i-1}, e_i)$

Embeddings
$e_i = x_i \cdot W_{ev}$

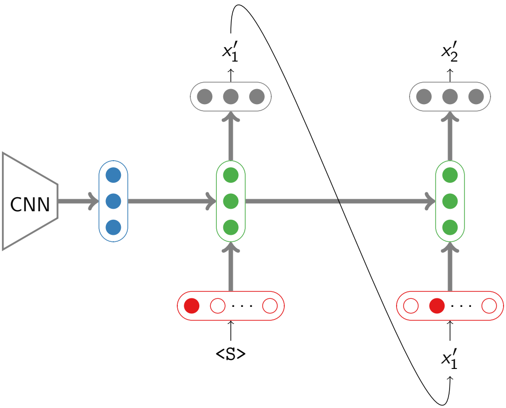
Input

Decoding with Multimodal Language Models



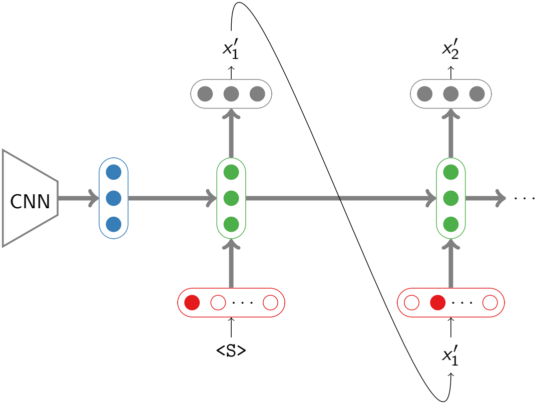
- Initialise with image features and <S> token

Decoding with Multimodal Language Models



- Initialise with image features and <S> token
- Feed sampled word x'_1 as input at the next timestep

Decoding with Multimodal Language Models



- Initialise with image features and <S> token
- Feed sampled word x_1' as input at the next timestep
- Decode until emit <E> token

Are we making progress?

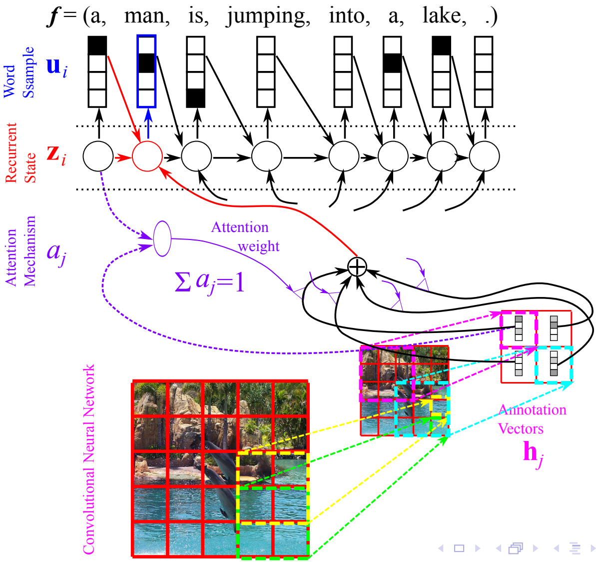
[Karpathy and Fei-Fei, 2015]

Girl in pink dress is jumping in air.

2011 - - - → 2012 - - - → 2013 - - - → 2014

Visual Attention

[Xu et al., 2015]



Are we making progress?
[Xu et al., 2015]

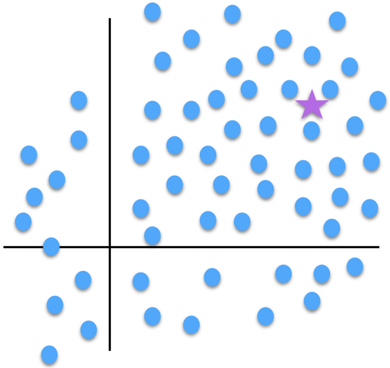
A woman is throwing a frisbee in a park.

2011 - - - → 2012 - - - → 2013 - - - → 2014 - - - → 2015

Nearest-neighbour approaches

[Devlin et al., 2015, Yagcioglu et al., 2015]

Do we even need to generate descriptions?

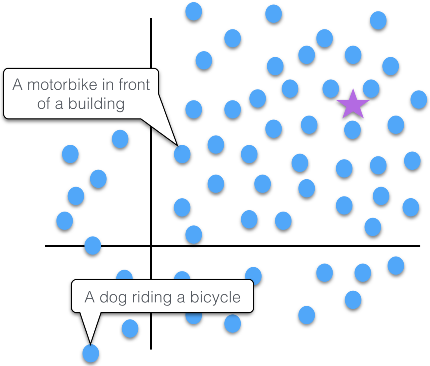


1. Visual similarity space: *cosine*(FC_7, FC_7)

Nearest-neighbour approaches

[Devlin et al., 2015, Yagcioglu et al., 2015]

Do we even need to generate descriptions?

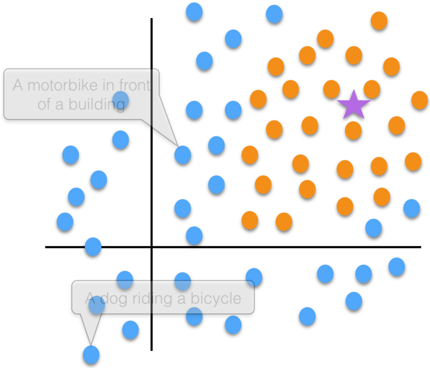


1. Visual similarity space: $\mathit{cosine}(\mathrm{FC}_7, \mathrm{FC}_7)$

Nearest-neighbour approaches

[Devlin et al., 2015, Yagcioglu et al., 2015]

Do we even need to generate descriptions?

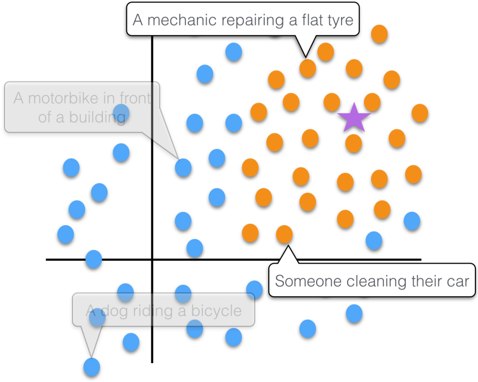


1. Visual similarity space: *cosine*(FC_7, FC_7)
2. Gather C captions of the K nearest neighbours

Nearest-neighbour approaches

[Devlin et al., 2015, Yagcioglu et al., 2015]

Do we even need to generate descriptions?

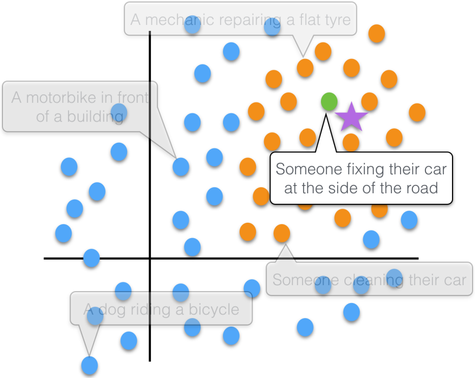


1. Visual similarity space: *cosine*(FC_7, FC_7)
2. Gather C captions of the K nearest neighbours

Nearest-neighbour approaches

[Devlin et al., 2015, Yagcioglu et al., 2015]

Do we even need to generate descriptions?



1. Visual similarity space: *cosine*(FC_7, FC_7)
2. Gather C captions of the K nearest neighbours
3. Retrieve the consensus caption $argmax_{c \in C} \sum_{c' \in C} sim(c, c')$ $sim(\cdot, \cdot)$ CIDEr or BLEU

Overview of Recent Approaches

CNN-RNN

[Vinyals et al., 2015]
[Karpathy and Fei-Fei, 2015]
[Donahue et al., 2015]
[Mao et al., 2015]

Deeper Networks

[Donahue et al., 2015]
[Mao et al., 2015]

Additional Evidence

[Jia et al., 2015]
[You et al., 2016]

Alternative LMs

[Kiros et al., 2014]
[Fang et al., 2015]

Retrieval Approaches

[Devlin et al., 2015]
[Yagcioglu et al., 2015]

Attention-based Models

[Xu et al., 2015]

Evaluating Descriptions

Hyp

A man is throwing his bike at a bus

Refs

A man putting a bike on the front of a bus

A young bicyclist is parking his bike on the bus rack

A man mounting his bike in the front of a city bus

A man and a bike by a large bus

A man is loading his bicycle on the front rack of a bus

$\boxed{\text{Compare}}$

42?

Current Approaches to Evaluation

Inspired by machine translation, we use:

- **BLEU** n-gram precision [Papineni et al., 2002]
- **ROUGE** skip-gram recall [Lin and Hovy, 2003]
- **Meteor** word/stem/synset/paraphrase matching [Denkowski and Lavie, 2014]

Current Approaches to Evaluation

Inspired by machine translation, we use:

BLEU	n-gram precision [Papineni et al., 2002]
ROUGE	skip-gram recall [Lin and Hovy, 2003]
Meteor	word/stem/synset/paraphrase matching [Denkowski and Lavie, 2014]

As a community, we developed:

Ranking	image–sentence retrieval & vice-versa [Hodosh et al., 2013]
CIDEr	consensus-based sentence similarity [Vedantam et al., 2015]

Current Approaches to Evaluation

Inspired by machine translation, we use:

BLEU	n-gram precision [Papineni et al., 2002]
ROUGE	skip-gram recall [Lin and Hovy, 2003]
Meteor	word/stem/synset/paraphrase matching [Denkowski and Lavie, 2014]

As a community, we developed:

Ranking	image–sentence retrieval & vice-versa [Hodosh et al., 2013]
CIDEr	consensus-based sentence similarity [Vedantam et al., 2015]

What does it mean when we outperform human–human agreement?

Moving forwards: Better evaluation measures

[Elliott and Keller, 2014, Vedantam et al., 2015]

New text-based similarity measures will be very broadly useful.
But we need larger *open* datasets of human judgements.

	Spearman's ρ
CIDEr	0.578
Meteor	0.524
ROUGE SU-4	0.435
BLEU-4	0.429
BLEU-1	0.345
TER	-0.279

Flickr8K, n=17,466, Likert-scale=1,...,4

Moving forwards: Back to human judgements

- Has no incorrect information
  [Mitchell et al., 2012]
- Is relevant for this image
  [Li et al., 2011, Yang et al., 2011]
- Is creatively constructed
  [Li et al., 2011]
- Is human-like
  [Mitchell et al., 2012]
- Is grammatically correct
  [Yang et al., 2011, Mitchell et al., 2012, Kuznetsova et al., 2012, Elliott and Keller, 2013, inter-alia]
- Accurately describes the image
  [Kulkarni et al., 2011, Li et al., 2011, Mitchell et al., 2012, Kuznetsova et al., 2012, Elliott and Keller, 2013]

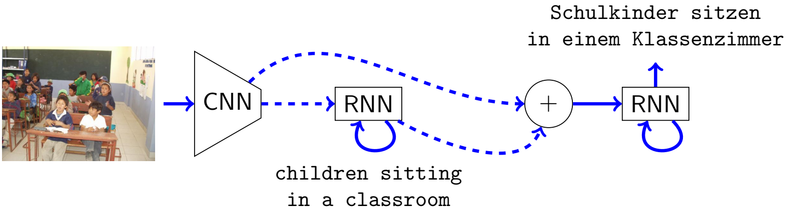
Next: Describing historic image collections

"Two people are walking down at river in a wooded area"

Full collection: `https://staff.fnwi.uva.nl/d.elliott/loc/`

Next: Image Description in Multiple Languages

[Elliott et al., 2015] [Hitschler et al., 2016] [Specia et al., 2016]



Next: Image Description in Multiple Languages

[Elliott et al., 2015] [Hitschler et al., 2016] [Specia et al., 2016]

English	a man is standing on a grey rock in the foreground ✗
German	bergsteiger klettern auf einen sehr steilen eishang ✓
Transfer	tourists are climbing up a snowy slope ✓

Resources

Survey: Automatic description generation from images: A survey of models, datasets, and evaluation measures. Bernardi et al. 2016. Journal of Artificial Intelligence Research.

NeuralTalk: `https://github.com/karpathy/neuraltalk2`

Arctic Captions: `https://github.com/kelvinxu/arctic-captions`

GroundedTranslation: `https://github.com/elliottd/GroundedTranslation`

Flickr30K: `http://shannon.cs.illinois.edu/DenotationGraph/`

MS COCO: `http://www.mscoco.org`

Multi30K: `http://www.statmt.org/wmt16/multimodal-task.html`

Visual Question Answering

Motivation

Image description:

- is a passive task
- users may not care about complete descriptions [Gao et al., 2015]
- descriptions add nothing to what a person has already perceived [Mostafazadeh et al., 2016]

Motivation

Image description:

- is a passive task
- users may not care about complete descriptions [Gao et al., 2015]
- descriptions add nothing to what a person has already perceived [Mostafazadeh et al., 2016]

Visual Question Answering:

- focus on specific aspects of language and vision
- multiple choice answers are easier to evaluate
  → easier to measure progress

Multiple-choice questions

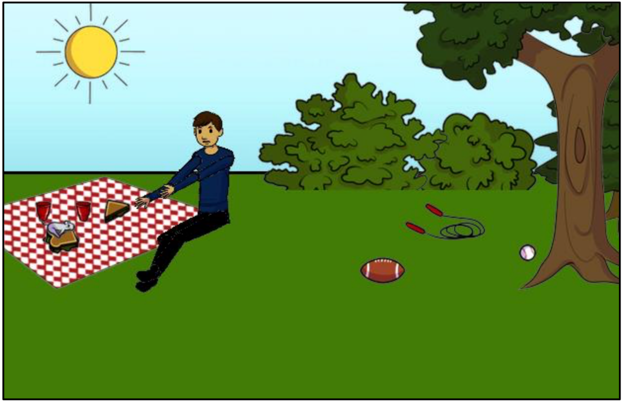
Visual7W: [Zhu et al., 2016]

Who has a hat on?

- A woman.
- A dog.
- A child.
- The man.

Multiple-choice questions

VQA: [Antol et al., 2015]



Is this person expecting company?
What is just under the tree?

Open-ended questions

FM-IQA: [Gao et al., 2015]

公共汽车是什么颜色的？
What is the color of the bus?

公共汽车是红色的。
The bus is red.

Open-ended questions

Visual Madlibs: [Yu et al., 2015]

Q: Describe what happened immediately after this picture was taken.

Open-ended questions

Visual Madlibs: [Yu et al., 2015]

Q: Describe what happened immediately after this picture was taken.

A: They drove around.

Datasets

	Images	Q-A Pairs	Open-ended	Multiple Choice
DAQUAR [Malinowski et al., 2015]	1,500	13,000	Yes	No
Visual QA [Antol et al., 2015]	250,000	760,000[2]	Yes	Yes
Visual Madlibs [Yu et al., 2015]	10,000	360,000[3]	Yes	Yes
Visual7W [Zhu et al., 2016]	47,000	330,000	Yes	Yes
COCO-QA [Ren et al., 2015]	124,000	118,000	Yes	Yes
FM-IQA [Gao et al., 2015]	150,000	310,000[4]	Yes	Yes

[2] 10M answers
[3] 12 question types
[4] Zh → En

Evaluation Methodologies

Your model proposes an answer A
There is one correct answer H (human)

- Accuracy:

  harsh with only one human reference.

  $H = orange \qquad A = mandarin$ ✗

- Wu-Palmer Similarity [Wu and Palmer, 1994]

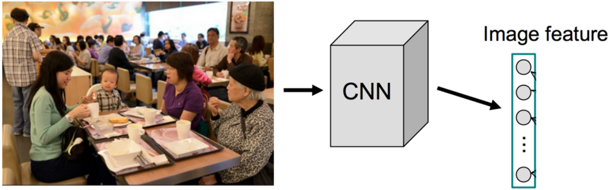
  $$\text{WUP}(x, y) = 2 * \frac{\text{depth of most specific common ancestor}}{\text{depth}(x) * \text{depth}(y)}$$

- Or collect many human answers (e.g. 10)

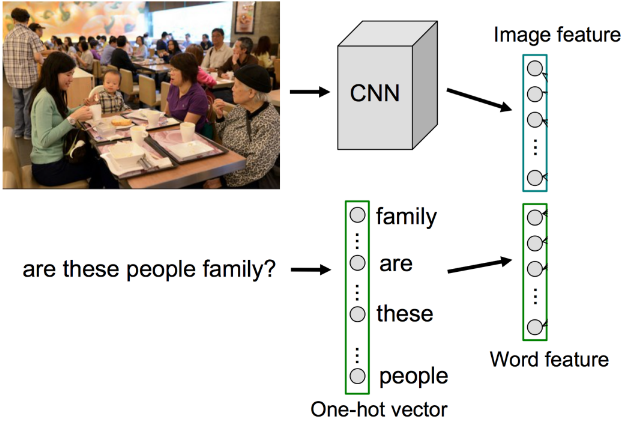
  $\text{Accuracy} = \min(\frac{A}{3}, 1)$

A Bag-of-Words Baseline

[Zhou et al., 2015]

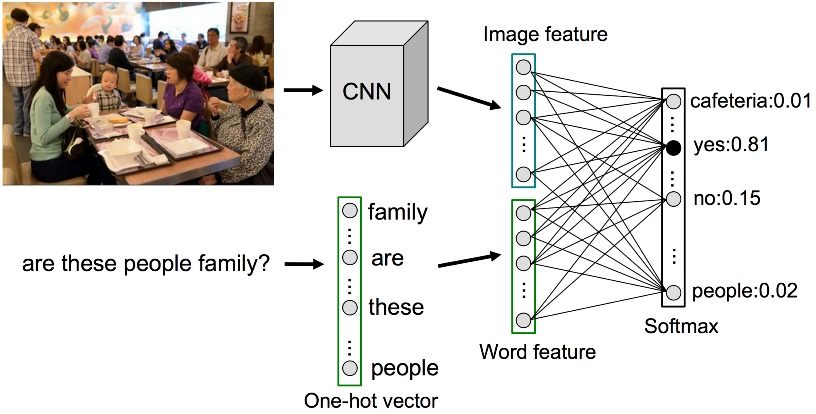


A Bag-of-Words Baseline
[Zhou et al., 2015]



A Bag-of-Words Baseline

[Zhou et al., 2015]



Another Baseline!
[Jabri et al., 2016]

SCORE

8192D MLP

300D | 2048D | 300D

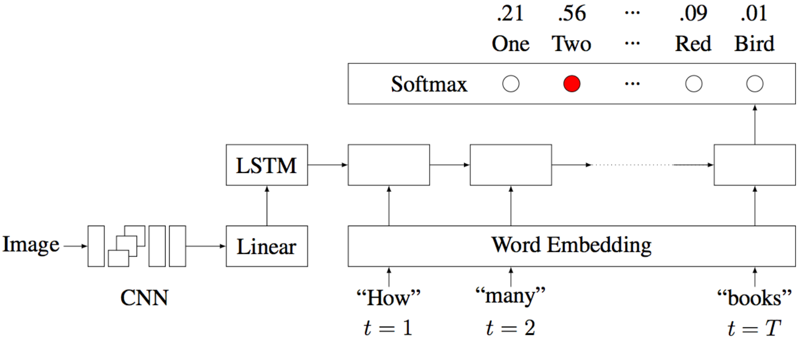
BoW | CNN | BoW

What is the weather like?

Sunny

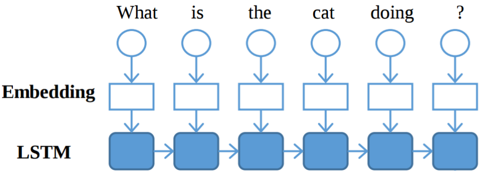
CNN-RNN for Multiple Choice VQA

[Ren et al., 2015]



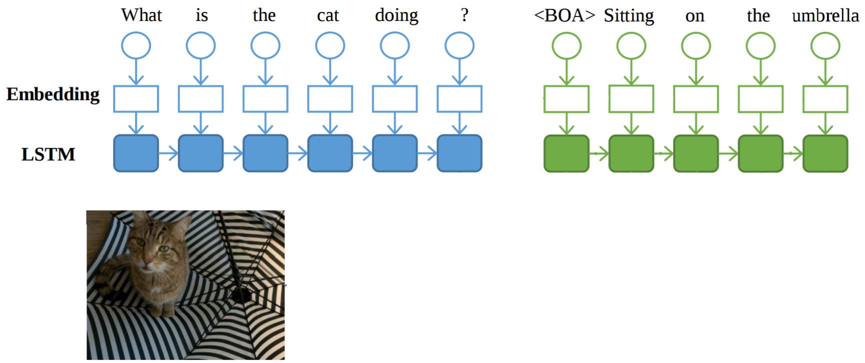
Multimodal Fusion and Answer Generation

[Gao et al., 2015]



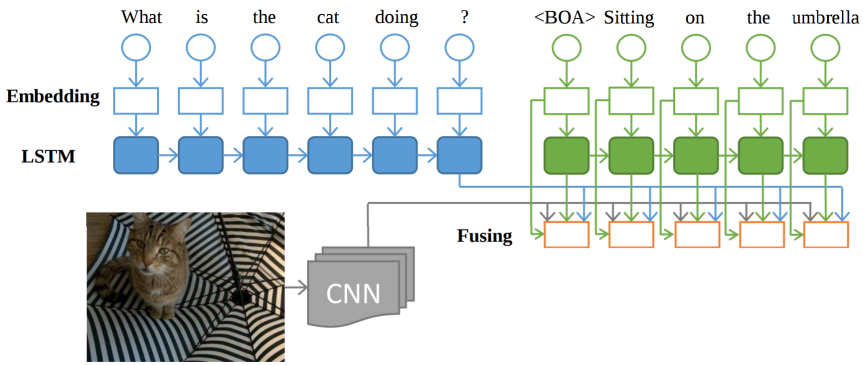
Multimodal Fusion and Answer Generation

[Gao et al., 2015]



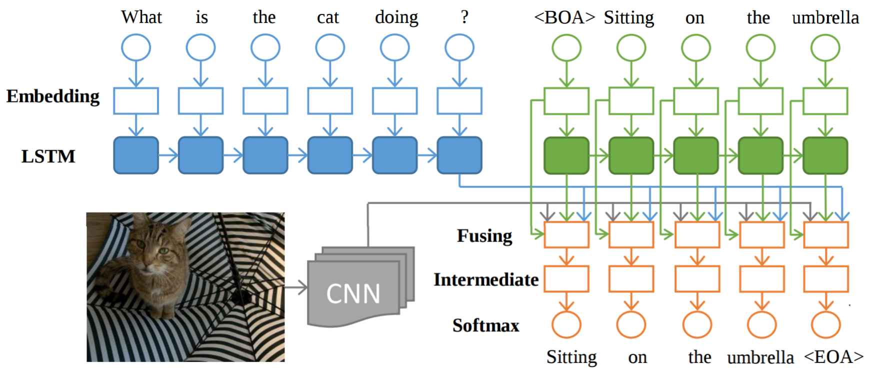
Multimodal Fusion and Answer Generation

[Gao et al., 2015]



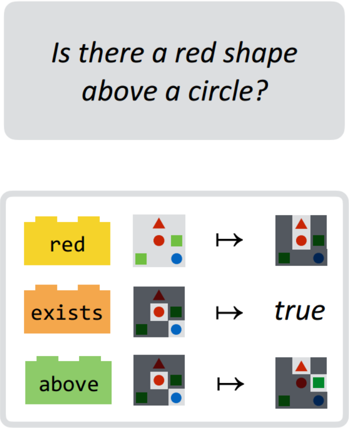
Multimodal Fusion and Answer Generation

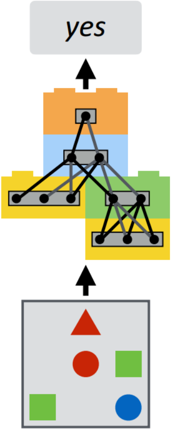
[Gao et al., 2015]

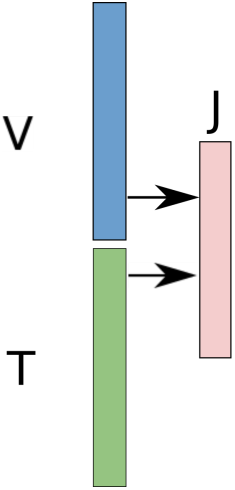


Composing Neural Networks for VQA

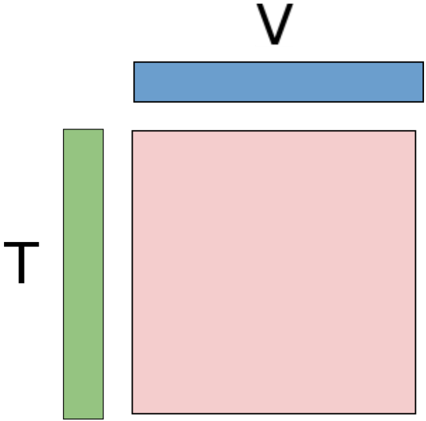
[Andreas et al., 2016]







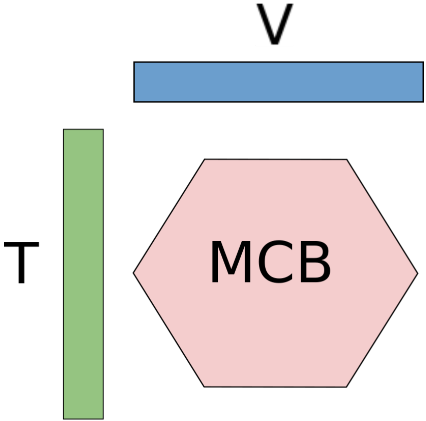
- Multimodal representations are typically a sum over projections from each modality
- $\mathrm{J} = \mathrm{W}_{jv} \cdot v + \mathrm{W}_{jt} \cdot t$
- Additive interaction between modalities ✗



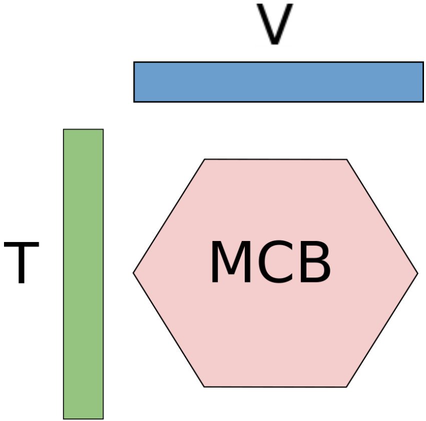
- Bilinear pooling allows for multiplicative interactions between vectors ✓
- BP $= \mathrm{v} \otimes \mathrm{t}$
- Too many parameters ✗

Multimodal Compact Bilinear Pooling

Compact Bilinear Pooling [Gao et al., 2016]



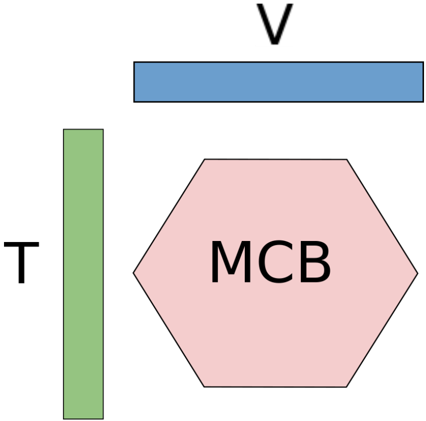
- Multiplicative interactions between vectors ✓
- Definable parameters ✓
- Count Sketch function Ψ



- Ψ: $x \in \mathbb{R}^n \rightarrow y \in \mathbb{R}^d$ $\quad$ d $\ll$ n

Multimodal Compact Bilinear Pooling

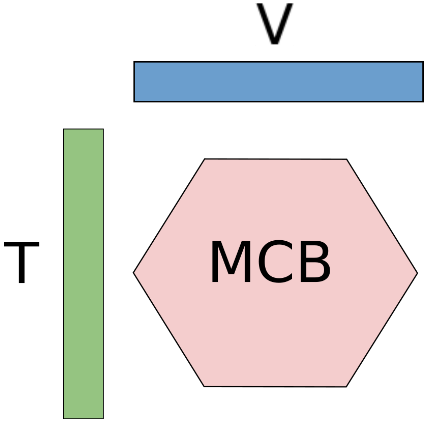
Compact Bilinear Pooling [Gao et al., 2016]



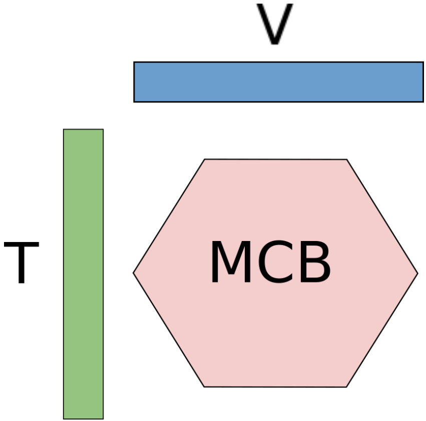
- Ψ: $x \in \mathbb{R}^n \rightarrow y \in \mathbb{R}^d$ $\quad d \ll n$
- $y = \Psi(x, h, s)$

Multimodal Compact Bilinear Pooling

Compact Bilinear Pooling [Gao et al., 2016]



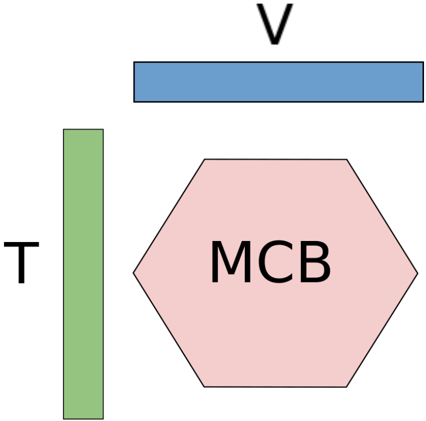
- Ψ: $x \in \mathbb{R}^n \rightarrow y \in \mathbb{R}^d$ $\quad$ d ≪ n
- $y = \Psi(x, h, s)$
- h: $x[i] \rightarrow y[j]$ $\quad$ randomly fixed



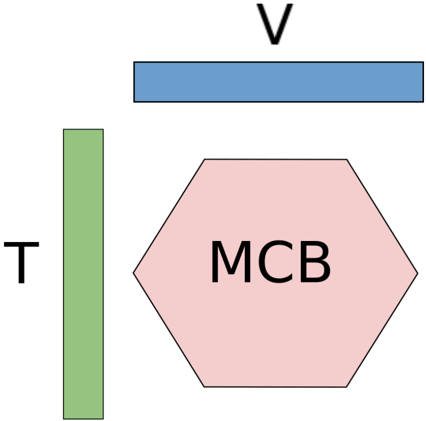
- Ψ: $x \in \mathbb{R}^n \rightarrow y \in \mathbb{R}^d$ $\quad d \ll n$
- $y = \Psi(x, h, s)$
- h: $x[i] \rightarrow y[j]$ randomly fixed
- s: $<s_1, \ldots, s_n>$ $\quad s_i \in \{-1, 1\}$

Multimodal Compact Bilinear Pooling

Compact Bilinear Pooling [Gao et al., 2016]



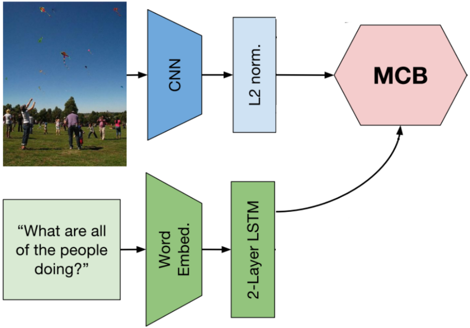
- Ψ: $x \in \mathbb{R}^n \rightarrow y \in \mathbb{R}^d$ $\quad$ $d \ll n$
- $y = \Psi(x, h, s)$
- h: $x[i] \rightarrow y[j]$ $\quad$ randomly fixed
- s: $<s_1, \ldots, s_n>$ $\quad$ $s_i \in \{-1, 1\}$
- $y[h_j] \leftarrow x_j \cdot s_j + y[h_j]$



- Ψ: $x \in \mathbb{R}^n \rightarrow y \in \mathbb{R}^d$ $\quad d \ll n$
- $y = \Psi(x, h, s)$
- h: $x[i] \rightarrow y[j]$ $\quad$ randomly fixed
- s: $<s_1, \ldots, s_n>$ $\quad s_i \in \{-1, 1\}$
- $y[h_j] \leftarrow x_j \cdot s_j + y[h_j]$
- $MCB = FFT^{-1}(FFT(v') \odot FFT(t'))$

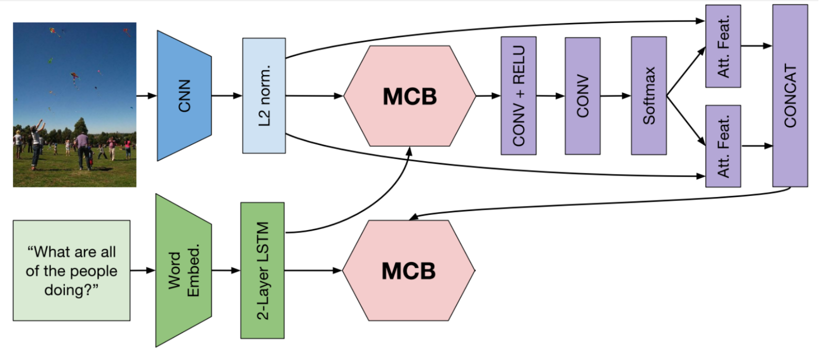
Multimodal Compact Bilinear Pooling for VQA

[Fukui et al., 2016]



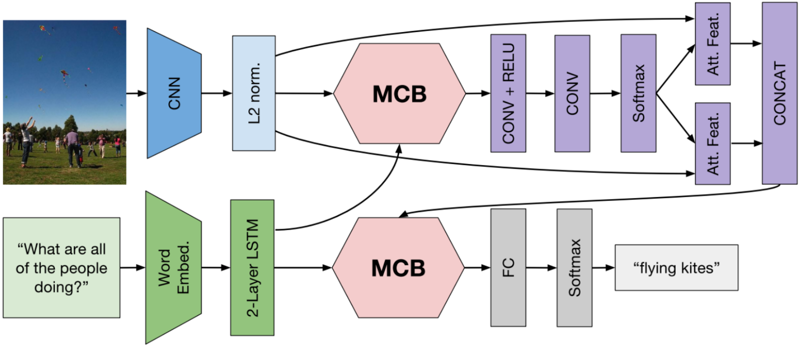
Multimodal Compact Bilinear Pooling for VQA

[Fukui et al., 2016]



Multimodal Compact Bilinear Pooling for VQA

[Fukui et al., 2016]



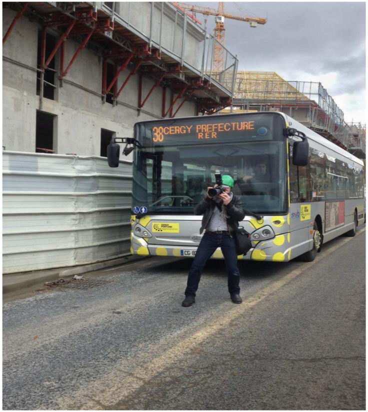
Are we making progress?
[Gao et al., 2015]

Q: *Is this guy playing tennis?*
A: Yes

Are we making progress?
[Ren et al., 2015]

Q: *What colour is the cat?*
A: Black

Are we making progress?
[Jabri et al., 2016]



Q: *What is behind the photographer?*
A: Bus

Are we making progress?

[Fukui et al., 2016]

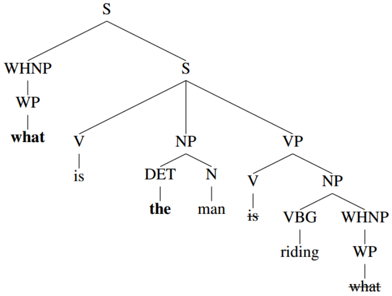
Q: *What moves people to the top of the hill?*

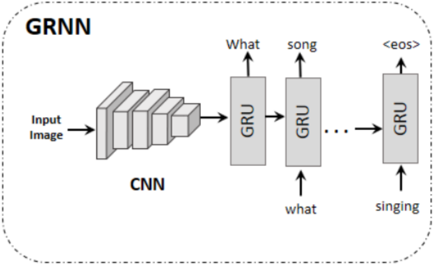
A: Ski lift

Next: Question Generation

[Ren et al., 2015, Mostafazadeh et al., 2016]

Learn how to ask questions about images





Resources

Survey Visual Question Answering: A Survey of Methods and Datasets. We et al. (2016). CoRR/1607.05910

MCB https://github.com/akirafukui/vqa-mcb/
NMN http://github.com/jacobandreas/nmn2
Neural-QA https://github.com/mateuszmalinowski/visual_turing_test-tutorial/

Visual7W http://web.stanford.edu/~yukez/visual7w/
VQA http://www.visualqa.org
FM-IQA http://idl.baidu.com/FM-IQA.html
DAQUAR http://www.mpi-inf.mpg.de/departments/computer-vision-and-multimodal-computing/research/vision-and-language/visual-turing-challenge/
Visual Madlibs http://tamaraberg.com/visualmadlibs/
COCO-QA http://www.cs.toronto.edu/~mren/imageqa/data/cocoqa/

More Multimodal Understanding: Video Description

[Thomason et al., 2014, Venugopalan et al., 2015]

↓

An grumpy old man is lecturing a kid

More: Visual Storytelling

[Huang et al., 2016]

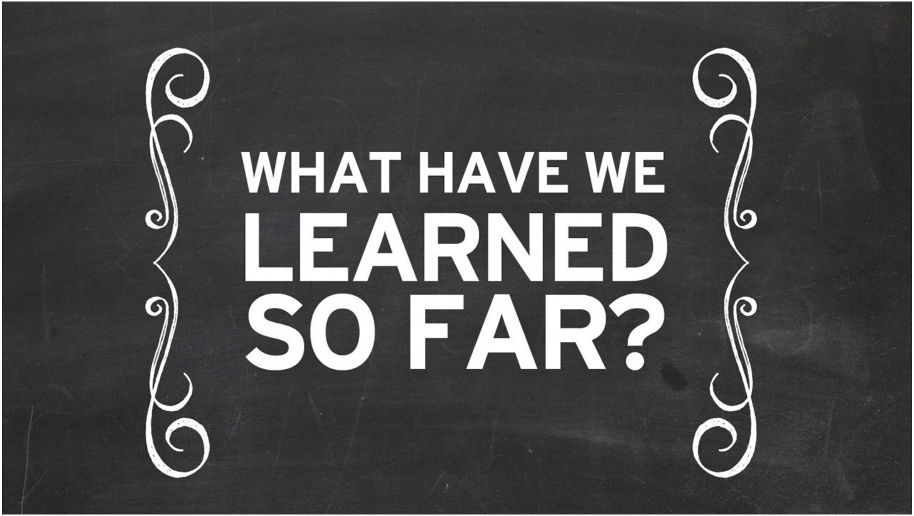
1 2 3

The dog was ready to go. | He had a great time on the hike. | And was very happy to be in the field.

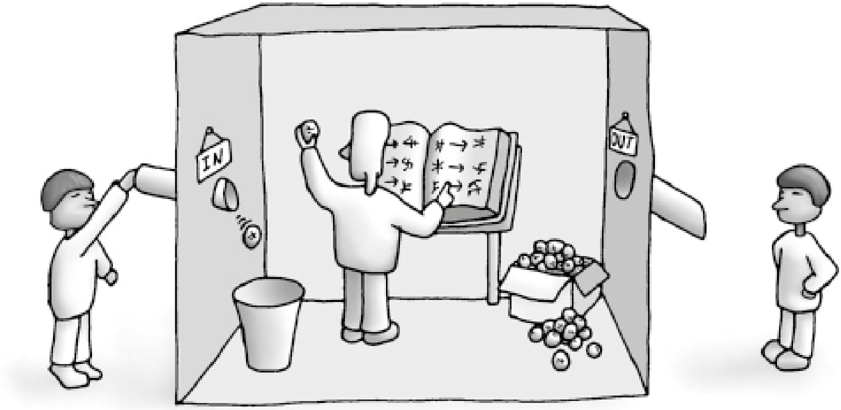
Photos by kameraschwein / CC BY-NC-ND 2.0

Final Words

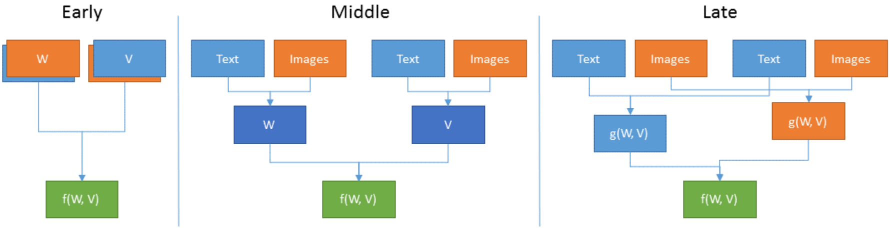
Recap



Grounding



Representational grounding: Multi-modal fusion



Referential grounding: Cross-modal mapping

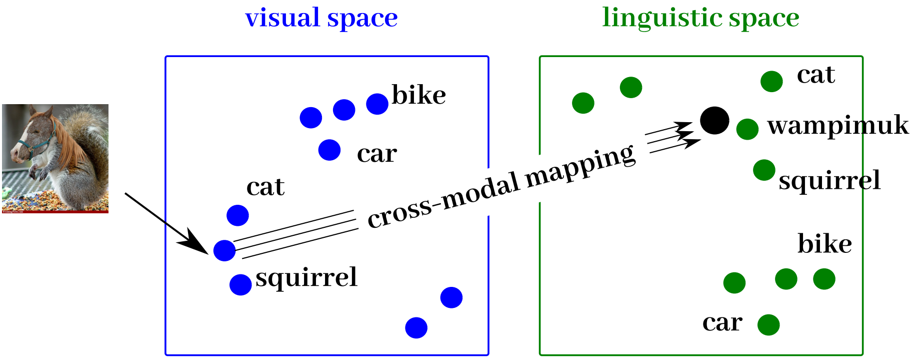
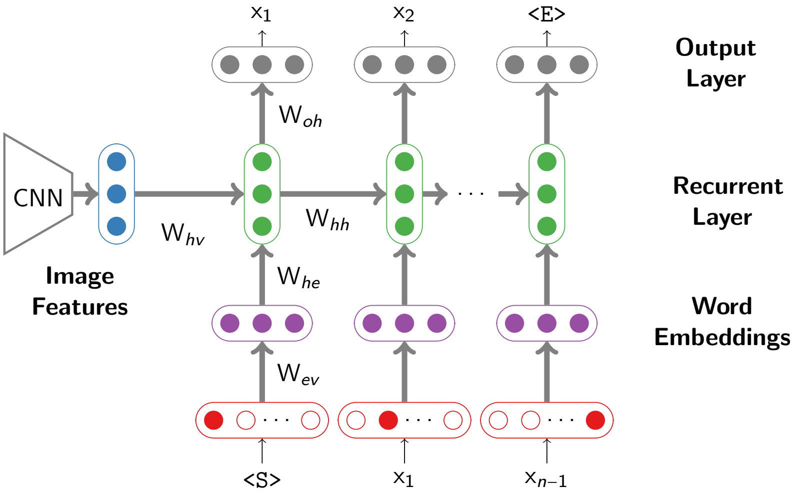
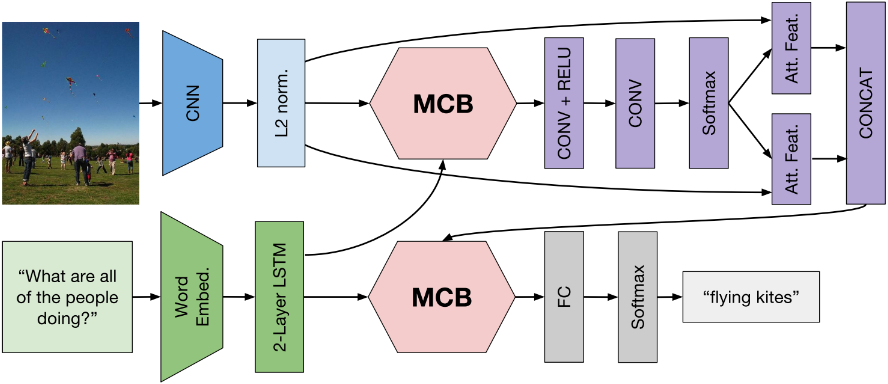


Image Description

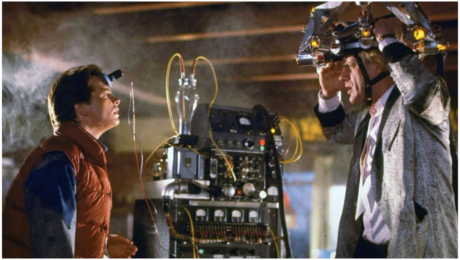


Visual question answering



The future of multi-modal NLP

- Going beyond vision
- Current issues & open problems
- Applications



Going beyond vision

However, if the objective is to ground semantic representations in perceptual information, **why stop at image data**? The meaning of *violin* is surely not only grounded in its visual properties, such as shape, color and texture, but also in its sound, pitch and timbre.

Other perceptual modalities

- Auditory grounding [Lopopolo and van Miltenburg, 2015, Kiela and Clark, 2015]
- Olfactory/gustatory grounding [Kiela et al., 2015a]
- Haptic.. ?
- Multi-modal has mostly been bi-modal so far, how about "poly-"modal.. ?
- Videos [Yu and Siskind, 2013, Regneri et al., 2013]
- Robotics [Coradeschi et al., 2013]

Current issues: Data

- A lot of unexploited unstructured data available
  - Movies, scripts, plays
  - Music, audiobooks
  - .. whatever else the Web has to offer
- Less supervision, but the data is there
- Think of ways to become less dependent on humans

Current issues: Measuring progress

- Issues with **metrics**:
  - Spearman, BLEU, METEOR, etc. are not very apt
  - Should we return to directly asking humans?
  - What happens when we beat human scores? What does that mean?
- Issues with **tasks / datasets**
  - Focus less on state-of-the-art and more on novelty and generality
  - Should we evaluate on two tasks and tune on only one?
  - Do we need more datasets, bigger datasets, or both?
- Issues with **approach**
  - Ask more "why"-questions: why does this work? why should we care? where does it fail and why?
  - Picking ripe **and rotten** cherries

Current issues: Cognitive plausibility and explainability

- Successful approaches are not necessarily cognitive plausible. Example: sequence to sequence.
- At the very least, we should try **not to make mistakes humans wouldn't make**

- **New EU law** will also create a "right to explanation," whereby a user can ask for an explanation of an algorithmic decision that was made about them [Goodman and Flaxman, 2016]

Open problems: Objective functions

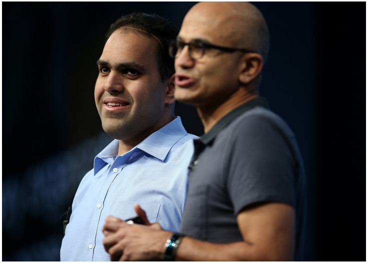
- Are we learning things the right way?
- For many applications, we will need **interactive learning**
- Should we learn by "utility" and start doing reinforcement everywhere?

Applications: Captioning other modalities

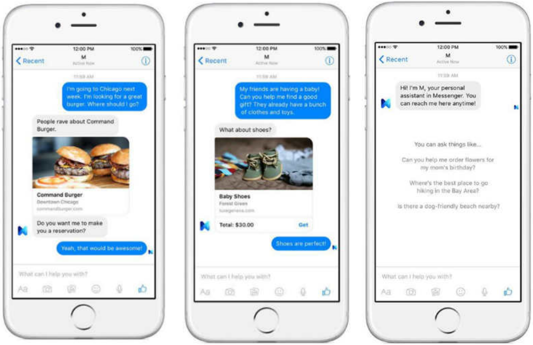
- Automatically describing audio
  - Describing **Chopin's étude Op. 25**: *One of the most stirring and most sublime pieces of music ever written: "Small-souled men, no matter how agile their fingers, should avoid it".* [Hofstadter, 1980]
- Digital vinologist / beerologist
  - Describing **Rochefort 10**: *The aroma is rich with dried fruit, such as figs, dates, and prunes. A light sourness is balanced by sweet molasses, followed by spice and pumpernickel bread.*

Applications: Audio descriptions of movies

- Automatically generating audio descriptions of movies
- Introducing scenes and dialogues in a smart way for visually-impaired
- Difficult problem: understanding the story, looking back, looking forward

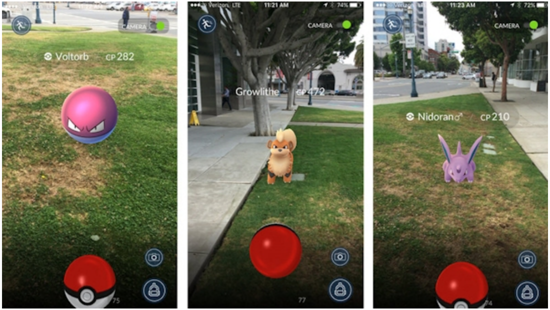


Applications: Clever assistants



- Next battlefield in industry: Cortana, Siri, Google Now, Facebook M
- Connecting modalities is essential

Applications: Virtual and augmented reality



- Understanding language relative to the environment
- Simultaneous perceptual and linguistic inputs

Applications: Video games

- Text-based games [Narasimhan et al., 2015]
- Learning to win by reading manuals [Branavan et al., 2011]
- Microsoft's Project Malmo

Learn more at ACL 2016

Mon 3E	16:50-17:10	Easy Things First: Installments Improve Referring Expression Generation for Objects in Photographs. Zarrieß and Schlangen
A	18:00-21:00	MMFeat: A Toolkit for Extracting Multimodal Features. Kiela
Tues 5A	13:40-13:56	The red one! On learning to refer to things based on discriminative properties. Lazaridou et al.
Tues 6E	15:30-17:10	Language and Vision Session
Wed 7E	10:10-10:30	Multimodal Pivots for Image Caption Translation. Hitschler, Schamoni and Riezler
Fri	09:00-17:30	5th Workshop on Vision & Language
WMT	09:20-09:45	A Shared Task on Multimodal MT and Crosslingual Image Description. Specia, Frank, Sima'an and Elliott
WMT	11:00-12:30	Poster Session on Multimodal Machine Translation and Cross-Lingual Image Description

References I

Andreas, J., Rohrbach, M., Darrell, T., and Klein, D. (2016).
Learning to compose neural networks for question answering.
In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1545–1554.

Andrews, M., Vigliocco, G., and Vinson, D. (2009).
Integrating experiential and distributional data to learn semantic representations.
Psychological review, 116(3):463.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., and Parikh, D. (2015).
Vqa: Visual question answering.
In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433.

Baroni, M., Dinu, G., and Kruszewski, G. (2014).
Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors.
In *Proceedings of ACL*, pages 238–247, Baltimore, MD.

Bergsma, S. and Goebel, R. (2011).
Using visual information to predict lexical preference.
In *Proceedings of RANLP*, pages 399–405.

Bergsma, S. and Van Durme, B. (2011).
Learning bilingual lexicons using the visual similarity of labeled web images.
In *IJCAI*, pages 1764–1769.

Berzak, Y., Barbu, A., Harari, D., Katz, B., and Ullman, S. (2015).
Do you see what i mean? visual resolution of linguistic ambiguities.
Conference on Empirical Methods in Natural Language Processing (EMNLP).

References II

Branavan, S., Silver, D., and Barzilay, R. (2011).
Learning to win by reading manuals in a Monte-Carlo framework.
In *Proceedings of ACL*, pages 268–277. Association for Computational Linguistics.

Bruni, E., Boleda, G., Baroni, M., and Tran, N. (2012).
Distributional semantics in technicolor.
In *ACL*, pages 136–145.

Bruni, E., Tran, G. B., and Baroni, M. (2011).
Distributional semantics from text and images.
In *Proceedings of the GEMS 2011 workshop on geometrical models of natural language semantics*, pages 22–32. Association for Computational Linguistics.

Bruni, E., Tran, N., and Baroni, M. (2014).
Multimodal distributional semantics.
Journal of Artifical Intelligence Research, 49:1–47.

Bulat, L., Kiela, D., and Clark, S. (2016).
Vision and feature norms: Improving automatic feature norm learning through cross-modal maps.
In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 579–588. Association for Computational Linguistics.

Chen, X., Fang, H., Lin, T., Vedantam, R., Gupta, S., Dollár, P., and Zitnick, C. L. (2015).
Microsoft COCO captions: Data collection and evaluation server.
CoRR, abs/1504.00325.

Chrupała, G., Kádár, A., and Alishahi, A. (2015).
Learning language through pictures.
page 112.

References III

Coradeschi, S., Loutfi, A., and Wrede, B. (2013).
A short review of symbol grounding in robotic and intelligent systems.
KI-Künstliche Intelligenz, 27(2):129–136.

Dale, R. and Reiter, E. (1995).
Computational interpretations of the gricean maxims in the generation of referring expressions.
Cognitive science, 19(2):233–263.

Denkowski, M. and Lavie, A. (2014).
Meteor Universal: Language Specific Translation Evaluation for Any Target Language.
In *EACL Workshop on Statistical Machine Translation*.

Devereux, B. J., Tyler, L. K., Geertzen, J., and Randall, B. (2014).
The centre for speech, language and the brain (cslb) concept property norms.
Behavior research methods, 46(4):1119–1127.

Devlin, J., Gupta, S., Girshick, R. B., Mitchell, M., and Zitnick, C. L. (2015).
Exploring nearest neighbor approaches for image captioning.
CoRR, abs/1505.04467.

Dinu, G., Lazaridou, A., and Baroni, M. (2015).
Improving zero-shot learning by mitigating the hubness problem.
In *Proceedings of ICLR Workshop Track*, San Diego, CA.

Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T. (2015).
Long-term recurrent convolutional networks for visual recognition and description.
In *CVPR*.

References IV

Elliott, D. and de Vries, A. P. (2015).
Describing images using inferred visual dependency representations.
In *ACL*, pages 42–52.

Elliott, D., Frank, S., and Hasler, E. (2015).
Multilingual image description with neural sequence models.
CoRR, abs/1510.04709.

Elliott, D., Frank, S., Sima'an, K., and Specia, L. (2016).
Multi30k: Multilingual english-german image descriptions.
In *Workshop on Vision and Language*.

Elliott, D. and Keller, F. (2013).
Image Description using Visual Dependency Representations.
In *EMNLP '13*, pages 1292–1302, Seattle, WA, U.S.A.

Elliott, D. and Keller, F. (2014).
Comparing Automatic Evaluation Measures for Image Description.
In *ACL*, pages 452–457.

Fang, H., Gupta, S., Iandola, F., Srivastava, R. K., Deng, L., Dollar, P., Gao, J., He, X., Mitchell, M., Platt, J. C., Lawrence Zitnick, C., and Zweig, G. (2015).
From captions to visual concepts and back.
In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., and Forsyth, D. (2010).
Every picture tells a story: Generating sentences from images.
In *ECCV*.

References V

Fazly, A., Alishahi, A., and Stevenson, S. (2010).
A probabilistic computational model of cross-situational word learning.
Cognitive Science, 34:1017–1063.

Felzenszwalb, P., McAllester, D., and Ramanan, D. (2008).
A discriminatively trained, multiscale, deformable part model.
In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE.

Feng, Y. and Lapata, M. (2010a).
How many words is a picture worth? automatic caption generation for news images.
In *Proceedings of the 48th annual meeting of the Association for Computational Linguistics*, pages 1239–1249. Association for Computational Linguistics.

Feng, Y. and Lapata, M. (2010b).
Visual information in semantic representation.
In *Proceedings of NAACL*, pages 91–99.

Frank, M. C., Tenenbaum, J. B., and Fernald, A. (2013).
Social and discourse contributions to the determination of reference in cross-situational word learning.
Language Learning and Development, 9(1):1–24.

Frome, A., Corrado, G., Shlens, J., Bengio, S., Dean, J., Ranzato, M., and Mikolov, T. (2013).
DeViSE: A deep visual-semantic embedding model.
In *Proceedings of NIPS*, pages 2121–2129, Lake Tahoe, NV.

Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., and Rohrbach, M. (2016).
Multimodal compact bilinear pooling for visual question answering and visual grounding.
CoRR, abs/1606.01847.

References VI

Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L., and Xu, W. (2015).
Are you talking to a machine? dataset and methods for multilingual image question.
In *Advances in Neural Information Processing Systems*, pages 2296–2304.

Gao, Y., Beijbom, O., Zhang, N., and Darrell, T. (2016).
In *Compact Bilinear Pooling*.

Goodman, B. and Flaxman, S. (2016).
European Union regulations on algorithmic decision-making and a "right to explanation".
ArXiv e-prints.

Grubinger, M., Clough, P., Müller, H., Deselaers, T., and Bank, W. (2006).
The IAPR TC-12 Benchmark : A New Evaluation Resource for Visual Information Systems.
In *OntoImage '06 at LREC '06*, Genoa, Italy.

Harnad, S. (1990).
The symbol grounding problem.
Physica D, 42:335–346.

He, K., Zhang, X., Ren, S., and Sun, J. (2015).
Deep residual learning for image recognition.
CoRR, abs/1512.03385.

Hill, F. and Korhonen, A. (2014).
Learning abstract concept embeddings from multi-modal data: Since you probably can't see what I mean.
In *Proceedings of EMNLP*, pages 255–265.

Hitschler, J., Schamoni, S., and Riezler, S. (2016).
Multimodal pivots for image caption translation.
In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2399–2409.

Hodosh, M., Young, P., and Hockenmaier, J. (2013).
Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics.
Journal of Artificial Intelligence Research, 47:853–899.

Hofstadter, D. R. (1980).
Gödel, escher, bach.
New Society.

Huang, T.-H. K., Ferraro, F., Mostafazadeh, N., Misra, I., Agrawal, A., Devlin, J., Girshick, R., He, X., Kohli, P., Batra, D., Zitnick, C. L., Parikh, D., Vanderwende, L., Galley, M., and Mitchell, M. (2016).
Visual storytelling.
In *NAACL*.

Jabri, A., Joulin, A., and van der Maaten, L. (2016).
Revisiting visual question answering baselines.
CoRR, abs/1606.08390.

Jackendoff, R. (2002).
Foundations of Language.
Oxford.

Jia, X., Gavves, E., Fernando, B., and Tuytelaars, T. (2015).
Guiding long-short term memory for image caption generation.
In *ICCV*.

References VIII

Johns, B. T. and Jones, M. N. (2012).
Perceptual inference through global lexical similarity.
Topics in Cognitive Science, 4(1):103–120.

Karpathy, A. and Fei-Fei, L. (2015).
Deep visual-semantic alignments for generating image descriptions.
In *CVPR*.

Kazemzadeh, S., Ordonez, V., Matten, M., and Berg, T. L. (2014).
Referitgame: Referring to objects in photographs of natural scenes.
In *EMNLP*, pages 787–798.

Kiela, D. and Bottou, L. (2014).
Learning image embeddings using convolutional neural networks for improved multi-modal semantics.
In *Proceedings of EMNLP*, pages 36–45.

Kiela, D., Bulat, L., and Clark, S. (2015a).
Grounding semantics in olfactory perception.
In *Proceedings of ACL*, pages 231–236, Beijing, China.

Kiela, D. and Clark, S. (2015).
Multi- and cross-modal semantics beyond vision: Grounding in auditory perception.
In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2461–2470, Lisbon, Portugal. Association for Computational Linguistics.

Kiela, D., Hill, F., and Clark, S. (2015b).
Specializing word embeddings for similarity or relatedness.
In *Proceedings of EMNLP*.

References IX

Kiela, D., Hill, F., Korhonen, A., and Clark, S. (2014).
Improving multi-modal representations using image dispersion: Why less is sometimes more.
In *Proceedings of ACL*, pages 835–841.

Kiela, D., Rimell, L., Vulić, I., and Clark, S. (2015c).
Exploiting image generality for lexical entailment detection.
In *Proceedings of ACL*, pages 119–124, Beijing, China. Association for Computational Linguistics.

Kiela, D., Vulić, I., and Clark, S. (2015d).
Visual bilingual lexicon induction with transferred convnet features.
In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 148–158, Lisbon, Portugal. Association for Computational Linguistics.

Kiros, R., Salakhutdinov, R., and Zemel, R. (2014).
Multimodal neural language models.
In *ICML*, pages 595–603.

Krahmer, E. and Van Deemter, K. (2012).
Computational generation of referring expressions: A survey.
Computational Linguistics, 38(1):173–218.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012a).
Imagenet classification with deep convolutional neural networks.
In Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012b).
ImageNet classification with deep convolutional neural networks.
In *Proceedings of NIPS*, pages 1106–1114.

References X

Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y. Y., Berg, A. C., and Berg, T. L. (2011).
Baby talk: Understanding and generating simple image descriptions.
In *CVPR*.

Kuznetsova, P., Ordonez, V., Berg, A. C., Berg, T. L., and Choi, Y. (2012).
Collective Generation of Natural Image Descriptions.
In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 359–368, Jeju Island, South Korea.

Kuznetsova, P., Ordonez, V., Choi, Y., and Berg, T. L. (2014).
TREETALK : Composition and Compression of Trees for Image Descriptions.
Transactions of the Association of Computational Linguistics, 2:351–362.

Lazaridou, A., Bruni, E., and Baroni, M. (2014).
Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world.
In *ACL (1)*, pages 1403–1414.

Lazaridou, A., Chrupała, G., Fernández, R., and Baroni, M. (2016).
Multimodal semantic learning from child-directed input.
In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 387–392.

Lazaridou, A., Dinu, G., and Baroni, M. (2015a).
Hubness and pollution: Delving into cross-space mapping for zero-shot learning.
In *Proceedings of ACL*, pages 270–280, Beijing, China.

Lazaridou, A., Nguyen, D. T., Bernardi, R., and Baroni, M. (2015b).
Unveiling the dreams of word embeddings: Towards language-driven image generation.
arXiv preprint arXiv:1506.03500.

References XI

Lazaridou, A., Pham, N. T., and Baroni, M. (2015c).
Combining language and vision with a multimodal skipgram model.
In *Proceedings of NAACL*.

LeCun, Y., Bengio, Y., and Hinton, G. (2015).
Deep learning.
Nature, 521(7553):436—444.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998).
Gradient-based learning applied to document recognition.
Proceedings of the IEEE, 86(11):2278–2324.

Leong, C. W. and Mihalcea, R. (2011a).
Going beyond text: A hybrid image-text approach for measuring word relatedness.
In *Proceedings of IJCNLP*, pages 1403–1407.

Leong, C. W. and Mihalcea, R. (2011b).
Measuring the semantic relatedness between words and images.
In *Proceedings of the Ninth International Conference on Computational Semantics*, pages 185–194. Association for Computational Linguistics.

Li, S., Kulkarni, G., Berg, T. L., Berg, A. C., and Choi, Y. Y. (2011).
Composing simple image descriptions using web-scale n-grams.
In *CoNLL*.

Lin, C.-Y. and Hovy, E. (2003).
Automatic evaluation of summaries using n-gram co-occurrence statistics.
In *NAA-CLHLT*.

References XII

Lopopolo, A. and van Miltenburg, E. (2015).
Sound-based distributional models.
In *Proceedings of the 11th International Conference on Computational Semantics (IWCS 2015)*.

Lowe, D. G. (2004).
Distinctive image features from scale-invariant keypoints.
International Journal of Computer Vision, 60(2):91–110.

Malinowski, M., Rohrbach, M., and Fritz, M. (2015).
Ask your neurons: A neural-based approach to answering questions about images.
In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–9.

Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., and Yuille, A. (2015).
Deep captioning with multimodal recurrent neural networks (m-RNN).
ICLR.

Mason, R. and Charniak, E. (2014).
Nonparametric method for data-driven image captioning.
In *Association for Computational Linguistics*, pages 592–598.

Matuszek, C., FitzGerald, N., Zettlemoyer, L., Bo, L., and Fox, D. (2012).
A Joint Model of Language and Perception for Grounded Attribute Learning.
In *Proc. of the 2012 International Conference on Machine Learning*, Edinburgh, Scotland.

McRae, K., Cree, G. S., Seidenberg, M. S., and McNorgan, C. (2005).
Semantic feature production norms for a large set of living and nonliving things.
Behavior Research Methods, 37(4):547–559.

References XIII

Mitchell, M., Han, X., Dodge, J., Mensch, A., Goyal, A., Berg, A. C., Yamaguchi, K., Berg, T. L., Stratos, K., Daume, III, H., and III (2012).
Midge: generating image descriptions from computer vision detections.
In *EACL*.

Mitchell, M., van Deemter, K., and Reiter, E. (2010).
Natural reference to objects in a visual domain.
In *Proceedings of the 6th international natural language generation conference*, pages 95–104. Association for Computational Linguistics.

Mohammad, S., Shutova, E., and Turney, P. (2016).
Metaphor as a medium for emotion: An empirical study.
Berlin, Germany.

Mostafazadeh, N., Misra, I., Devlin, J., Zitnick, L., Mitchell, M., He, X., and Vanderwende, L. (2016).
Generating natural questions about an image.
In *ACL*.

Narasimhan, K., Kulkarni, T. D., and Barzilay, R. (2015).
Language understanding for textbased games using deep reinforcement learning.
In *Proceedings of EMNLP*.

Nelson, D. L., McEvoy, C. L., , and Schreiber, T. A. (2004).
The University of South Florida free association, rhyme, and word fragment norms.
Behavior Research Methods, 36(3):402–407.

Norouzi, M., Mikolov, T., Bengio, S., Singer, Y., Shlens, J., Frome, A., Corrado, G. S., and Dean, J. (2014).
Zero-shot learning by convex combination of semantic embeddings.
In *Proceedings of ICLR*.

References XIV

Ordonez, V., Kulkarni, G., and Berg, T. L. (2011).
Im2text: Describing images using 1 million captioned photographs.
In *NIPS*.

Ortiz, G. M. L., Wolff, C., and Lapata, M. (2015).
Learning to interpret and describe abstract scenes.
In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1505–1515.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002).
BLEU: A method for automatic evaluation of machine translation.
In *ACL*.

Radovanović, M., Nanopoulos, A., and Ivanović, M. (2010).
Hubs in space: Popular nearest neighbors in high-dimensional data.
Journal of Machine Learning Research, 11:2487–2531.

Ramanathan, V., Joulin, A., Liang, P., and Fei-Fei, L. (2014).
Linking people with "their" names using coreference resolution.
In *European Conference on Computer Vision (ECCV)*.

Rashtchian, C., Young, P., Hodosh, M., and Hockenmaier, J. (2010).
Collecting image annotations using Amazon's Mechanical Turk.
In *AMT at NAACL '10*, pages 139–147, Los Angeles, CA, U.S.A.

Razavian, A. S., Azizpour, H., Sullivan, J., and Carlsson, S. (2014).
Cnn features off-the-shelf: an astounding baseline for recognition.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 806–813.

References XV

Regneri, M., Rohrbach, M., Wetzel, D., Thater, S., Schiele, B., and Pinkal, M. (2013).
Grounding action descriptions in videos.
Transactions of the Association for Computational Linguistics, 1:25–36.

Ren, M., Kiros, R., and Zemel, R. (2015).
Exploring models and data for image question answering.
In *Advances in Neural Information Processing Systems*, pages 2953–2961.

Roller, S. and Schulte im Walde, S. (2013).
A multimodal LDA model integrating textual, cognitive and visual modalities.
In *Proceedings of EMNLP*, pages 1146–1157.

Roy, D. (2002).
Learning visually grounded words and syntax of natural spoken language.
Evolution of communication, 4(1):33–56.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015).
ImageNet Large Scale Visual Recognition Challenge.
International Journal of Computer Vision (IJCV), 115(3):211–252.

Schlangen, D., Zarriess, S., and Kennington, C. (2015).
Resolving references to objects in photographs using the words-as-classifiers model.
arXiv preprint arXiv:1510.02125.

Searle, J. R. (1980).
Minds, brains, and programs.
Behavioral and brain sciences, 3(03):417–424.

References XVI

Shigeto, Y., Suzuki, I., Hara, K., Shimbo, M., and Matsumoto, Y. (2015).
Ridge regression, hubness, and zero-shot learning.
In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 135–151. Springer.

Shutova, E., Kiela, D., and Maillard, J. (2016).
Black holes and white rabbits: Metaphor identification with visual features.
In *Proceedings of NAACL-HTL 2016*, San Diego. Association for Computational Linguistics.

Silberer, C. and Lapata, M. (2012).
Grounded models of semantic representation.
In *Proceedings of EMNLP*, pages 1423–1433.

Silberer, C. and Lapata, M. (2014).
Learning grounded meaning representations with autoencoders.
In *Proceedings of ACL*, pages 721–732.

Simonyan, K. and Zisserman, A. (2015).
Very deep convolutional networks for large-scale image recognition.
In *ICLR '15*.

Siskind, J. (1996).
A computational study of cross-situational techniques for learning word-to-meaning mappings.
Cognition, 61:39–91.

Sivic, J. and Zisserman, A. (2003).
Video google: A text retrieval approach to object matching in videos.
In *Proceedings of ICCV*, pages 1470–1477.

References XVII

Socher, R., Ganjoo, M., Manning, C. D., and Ng, A. Y. (2013).
Zero shot learning through cross-modal transfer.
In *Advances in Neural Information Processing Systems 26*.

Specia, L., Frank, S., Sima'an, K., and Elliott, D. (2016).
A shared task on multimodal machine translation and crosslingual image description.
In *Proceedings of the First Conference on Machine Translation*, pages 543–553.

Srivastava, N. and Salakhutdinov, R. R. (2012).
Multimodal learning with deep boltzmann machines.
In *Advances in Neural Information Processing Systems 25*, pages 2222–2230.

Thomason, J., Venugopalan, S., Guadarrama, S., Saenko, K., and Mooney, R. (2014).
Integrating Language and Vision to Generate Natural Language Descriptions of Videos in the Wild.
In *COLING*.

Tsvetkov, Y., Boytsov, L., Gershman, A., Nyberg, E., and Dyer, C. (2014).
Metaphor detection with cross-lingual model transfer.
In *Proceedings of ACL*, Baltimore, MA.

Vedantam, R., Lawrence Zitnick, C., and Parikh, D. (2015).
Cider: Consensus-based image description evaluation.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575.

Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., and Saenko, K. (2015).
Sequence to sequence – video to text.
In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015).
Show and tell: A neural image caption generator.
In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Vulić, I., Kiela, D., Clark, S., and Moens, M. (2016).
Multi-modal representations for improved bilingual lexicon learning.
In *Proceedings of ACL*.

Wang, K., Yin, Q., Wang, W., Wu, S., and Wang, L. (2016).
A comprehensive survey on cross-modal retrieval.
arXiv preprint arXiv:1607.06215.

Weston, J., Bengio, S., and Usunier, N. (2011).
Wsabie: Scaling up to large vocabulary image annotation.
In *Proceedings of IJCAI*, pages 2764–2770.

Wu, Z. and Palmer, M. (1994).
Verbs semantics and lexical selection.
In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A. C., Salakhutdinov, R., Zemel, R. S., and Bengio, Y. (2015).
Show, attend and tell: Neural image caption generation with visual attention.
In *ICML*.

Yagcioglu, S., Erdem, E., Erdem, A., and Çakıcı Ruket (2015).
A distributed representation based query expansion approach for image captioning.
In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and The 7th International Joint Conference of the Asian Federation of Natural Language Processing*. ACL.

References XIX

Yang, Y., Teo, C. L., Daumé III, H., and Aloimonos, Y. (2011).
Corpus-guided sentence generation of natural images.
In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 444–454. Association for Computational Linguistics.

Yatskar, M., Vanderwende, L., and Zettlemoyer, L. (2014).
See no evil, say no evil: Description generation from densely labeled images.
SEM.

You, Q., Jin, H., Wang, Z., Fang, C., and Luo, J. (2016).
Image captioning with semantic attention.
In *CVPR*.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014).
From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions.
Transactions of the Association for Computational Linguistics.

Yu, C. and Ballard, D. (2004).
"a multimodal learning interface for grounding spoken language in sensory perceptions.
ACM Transactions on Applied Perception, 1:57–80.

Yu, H. and Siskind, J. M. (2013).
Grounded language learning from video described with sentences.
In *Proceedings of ACL 2013*, Sofia, Bulgaria.

Yu, L., Park, E., Berg, A. C., and Berg, T. L. (2015).
Visual madlibs: Fill in the blank image generation and question answering.
In *ICCV*.

References XX

Zeiler, M. D. and Fergus, R. (2014).
Visualizing and understanding convolutional networks.
In *Computer Vision–ECCV 2014*, pages 818–833. Springer.

Zhou, B., Tian, Y., Sukhbaatar, S., Szlam, A., and Fergus, R. (2015).
Simple baseline for visual question answering.
CoRR, abs/1512.02167.

Zhu, Y., Groth, O., Bernstein, M. S., and Fei-Fei, L. (2016).
Visual7w: Grounded question answering in images.
CVPR.

Zitnick, C. L. and Parikh, D. (2013).
Bringing semantics into focus using visual abstraction.
In *CVPR '13*.