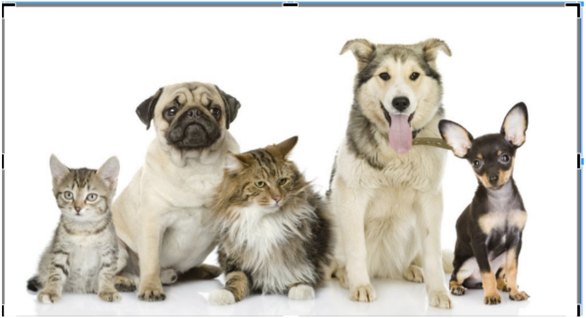
Learning quantities from vision and language

Raffaella Bernardi

University of Trento

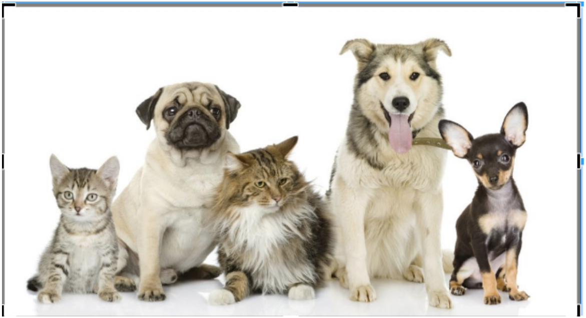
March 23, 2017

Cardinals and Quantifiers



Three of the animals are dogs. vs. *Most* of the animals are dogs

Cardinals and Quantifiers



Three of the animals are dogs. vs. *Most* of the animals are dogs

Quantifiers: are they in a scale?

Expected abstract scale: $<$no, few, some, most, all $>$

Q. How do we learn they are in this order?

Q. Do we take this order into account when using them?

Quantifiers: are they in a scale?

Expected abstract scale: $<$no, few, some, most, all $>$

Q. How do we learn they are in this order?

Q. Do we take this order into account when using them?

Quantifiers: are they in a scale?

Expected abstract scale: $<$no, few, some, most, all $>$

Q. How do we learn they are in this order?

Q. Do we take this order into account when using them?

Litteral vs. Pragmatic meaning

What do we learn from language, what from vision, what from both?

- **Conjecture 1:** we can learn their *litteral meaning* (respecting the abstract scale) from *images*.
- **Conjecture 2:** they can be represented by a *cross-modal function*.
- **Conjecture 3:** text corpora could help learning their *use*.

Litteral vs. Pragmatic meaning

What do we learn from language, what from vision, what from both?

- **Conjecture 1:** we can learn their *litteral meaning* (respecting the abstract scale) from *images*.
- **Conjecture 2:** they can be represented by a *cross-modal function*.
- **Conjecture 3:** text corpora could help learning their *use*.

Litteral vs. Pragmatic meaning

What do we learn from language, what from vision, what from both?

- **Conjecture 1:** we can learn their *litteral meaning* (respecting the abstract scale) from *images*.
- **Conjecture 2:** they can be represented by a *cross-modal function*.
- **Conjecture 3:** text corpora could help learning their *use*.

New Challenge for CV

From content words to Function words

- Most tasks considered so far involve processing of objects and lexicalised relations amongst objects (*content words*).
- Humans (even pre-school children) can abstract over raw data to perform certain types of higher-level reasoning, expressed in natural language by *function words*.

Operations inolved in quatifying

A logical strategy

Quantifiers require:

1. an approximate number estimation mechanism, acting over the relevant sets in the image;
2. a quantification comparison step.

A "logical" strategy:

1. from raw data to abstract set representation
2. from the latter to quantifiers.

Operations inolved in quatifying

A logical strategy

Quantifiers require:

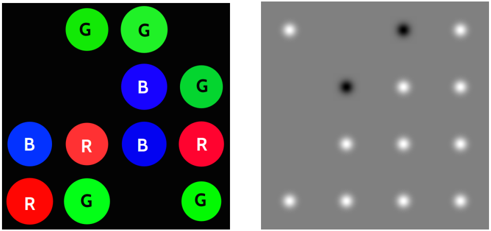
1. an approximate number estimation mechanism, acting over the relevant sets in the image;
2. a quantification comparison step.

A “logical” strategy:

1. from raw data to abstract set representation
2. from the latter to quantifiers.

Comparison step

Look, some green circles!: Learning to quantify from images (Sorodoc et al., 2016):

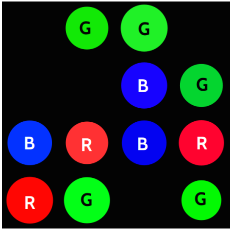


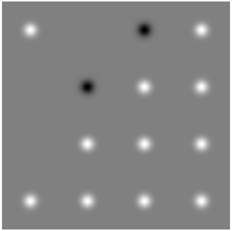
Very high results: NNs should be able to learn the second subtask quite easily.

Is the "logical" strategy a good one?

Comparison step

Look, some green circles!: Learning to quantify from images (Sorodoc et al., 2016):





Very high results: NNs should be able to learn the second subtask quite easily.
Is the "logical" strategy a good one?

Layout

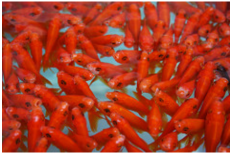
1. Learning quantification from images

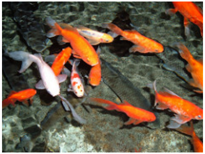
2. Quantifiers vs. Cardinals

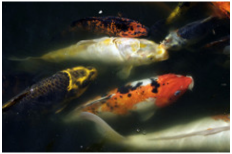
3. Behavioral Study

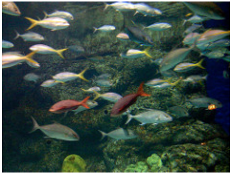
Learning quantification from images

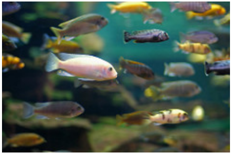
Pay attention to those sets! Learning quantification from images Sorodoc et. al. just submitted.

(a) 

(b) 

(c) 

(d) 

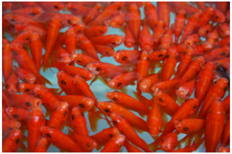
(e) 

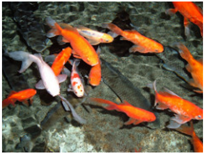
Query: ___ fish are red.

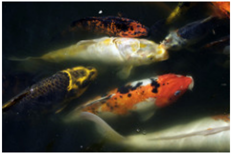
Answers: (a) All, (b) Most, (c) Some, (d) Few, (e) No

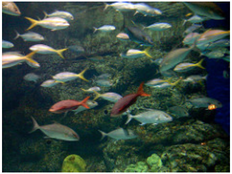
Learning quantification from images

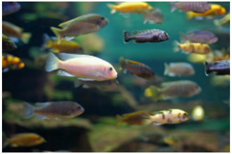
Pay attention to those sets! Learning quantification from images Sorodoc et. al. just submitted.

(a) 

(b) 

(c) 

(d) 

(e) 

Query: ___ fish are red.

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

Not raw data: All sorts of variances in place

The system cannot memorize correlations between

- type of objects and quantifiers
- property of objects and quantifiers
- number of objects and quantifiers

Quite challenging!

Quantifiers as proportions

$$Q \text{ of the } \underbrace{\textit{fish}}_{\textit{restrictor}} \text{ are } \underbrace{\textit{red}}_{\textit{scope}}.$$

We take quantifiers to be a fiexed relation:

$$\frac{|scope \cap restrictor|}{|restrictor|} \quad (e.g. \frac{|red \cap fish|}{|fish|})$$

Prevalence estimates (Khemlain et al 2009):

- No: 0%
- Few: 1% - 17% (inc)
- Some: 17 % - 70%
- Most: 70% (inc) - 99% (inc.)
- All: 100%

Quantifiers as proportions

$$Q \text{ of the } \underbrace{\textit{fish}}_{\textit{restrictor}} \text{ are } \underbrace{\textit{red}}_{\textit{scope}}.$$

We take quantifiers to be a fiexed relation:

$$\frac{|scope \cap restrictor|}{|restrictor|} \quad (e.g. \frac{|red \cap fish|}{|fish|})$$

Prevalence estimates (Khemlain et al 2009):

- No: 0%
- Few: 1% - 17% (inc)
- Some: 17 % - 70%
- Most: 70% (inc) - 99% (inc.)
- All: 100%

Quantifiers as proportions

$$Q \text{ of the } \underbrace{\textit{fish}}_{\textit{restrictor}} \text{ are } \underbrace{\textit{red}}_{\textit{scope}}.$$

We take quantifiers to be a fiexed relation:

$$\frac{|scope \cap restrictor|}{|restrictor|} \quad (e.g. \frac{|red \cap fish|}{|fish|})$$

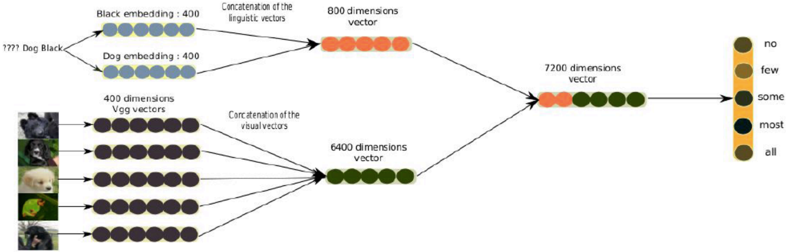
Prevalence estimates (Khemlain et al 2009):

- No: 0%
- Few: 1% - 17% (inc)
- Some: 17 % - 70%
- Most: 70% (inc) - 99% (inc.)
- All: 100%

Computer Vision Models

Start simple: concatenation. CNN+BOW

Zhou et al. *Simple Baseline for Visual Question Answering* 2015 (iBOWIMG)



Memorize correlations, no higher level abstraction

Computer Vision Models

Lesson learned from SoA: Memory and Attention

Memory process new information based on previous ones. (LSTM, GRU)
Attention Mechanism Use language to help making the representation of the image more focused
Stacked Attention use language to focus the visual representation and use the later to focus the linguistic representation.

Computer Vision Models

Lesson learned from SoA: Memory and Attention

Memory process new information based on previous ones. (LSTM, GRU)
Attention Mechanism Use language to help making the representation of the image more focused
Stacked Attention use language to focus the visual representation and use the later to focus the linguistic representation.

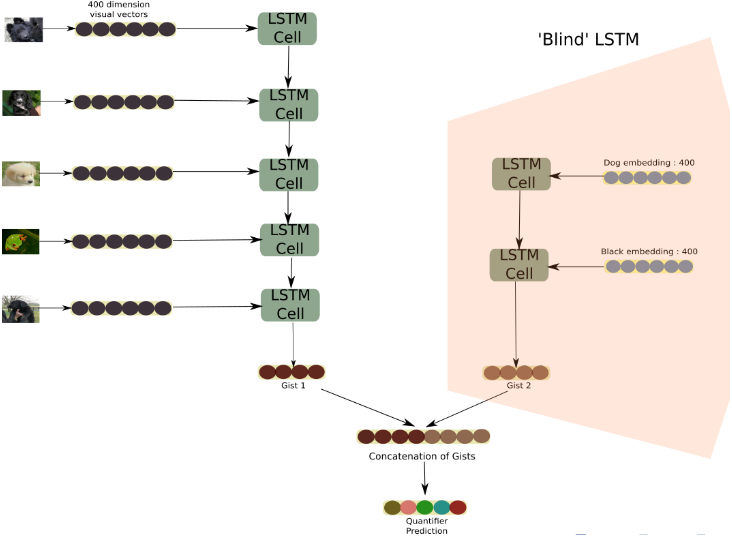
Computer Vision Models

Lesson learned from SoA: Memory and Attention

Memory process new information based on previous ones. (LSTM, GRU)
Attention Mechanism Use language to help making the representation of the image more focused
Stacked Attention use language to focus the visual representation and use the later to focus the linguistic representation.

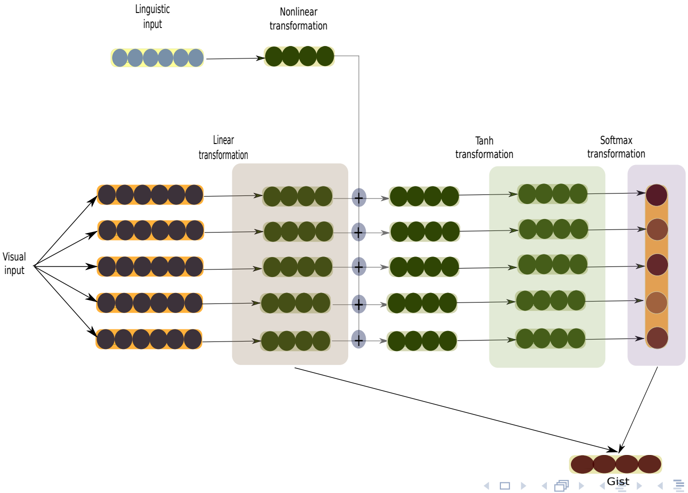
Sequential Processing

CNN+LSTM model



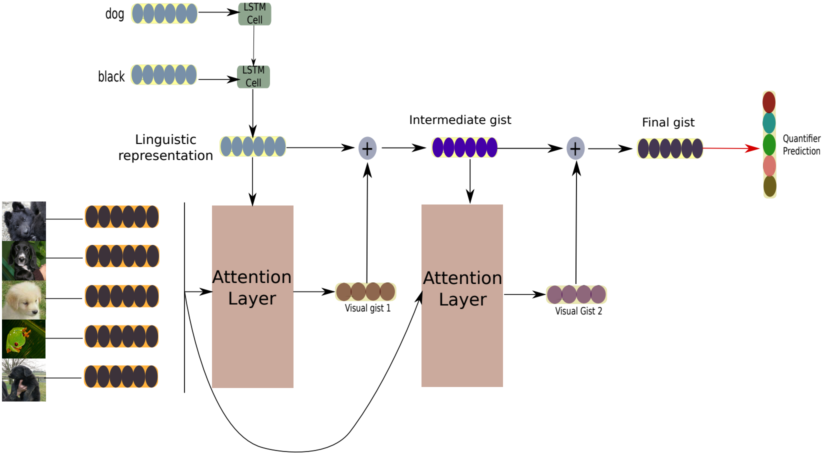
Attention Mechanism: SAN's attention layer

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



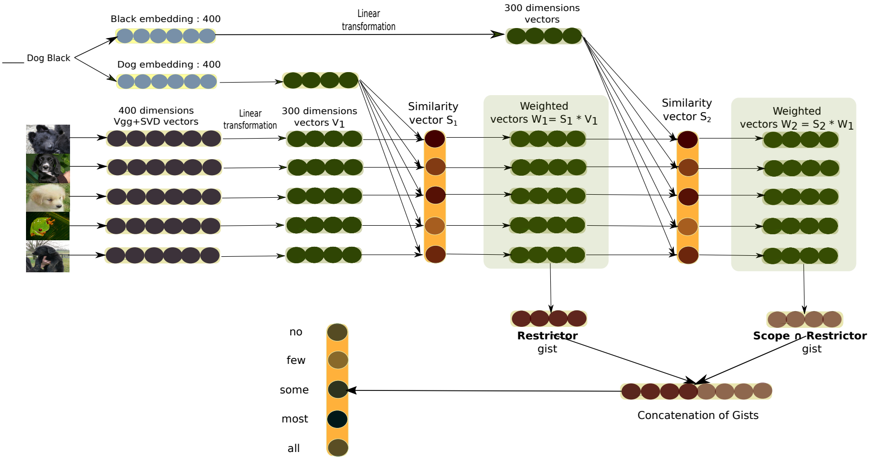
Stacked Attention Model

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



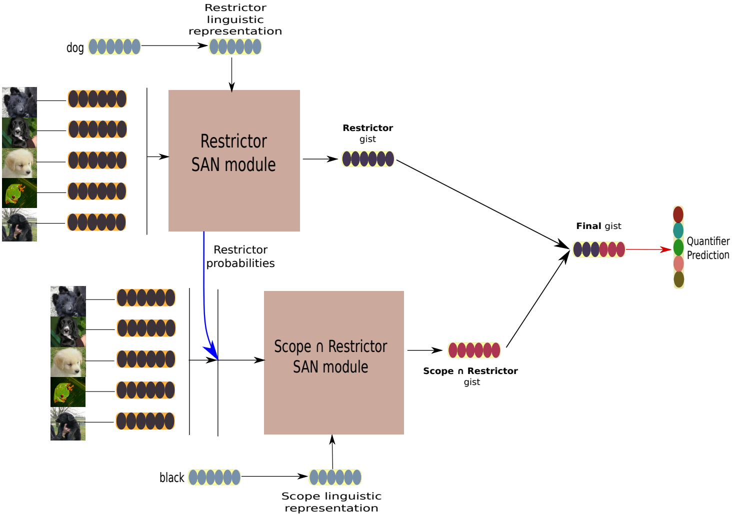
Linguistically motivated NNs with attention

Q Memory Network



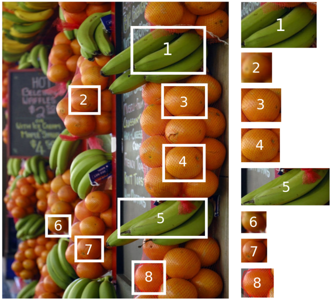
Linguistically motivated NNs with stacked attention

QSAN



Datasets: Q-COCO

ORIGINAL IMAGE **SCENARIO** **ANNOTATION**

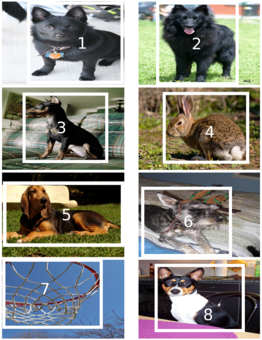


SCENARIO	ANNOTATION
1	**banana:** healthy
2	**orange:** fresh, tasty/delicious
3	**orange:** healthy, tasty/delicious, appetizing, fresh, round
4	**orange:** tasty/delicious, appetizing, fresh, cooked
5	**banana:** laying, healthy, tasty/delicious, horizontal, fresh, whole
6	**orange:** laying, round, fresh, appetizing tasty/delicious, whole, healthy
7	**orange:** tasty/delicious, fresh
8	**orange:** fresh

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

Datasets: Q-ImageNet

ORIGINAL IMAGE



SCENARIO

ANNOTATION

dog: furry, black

dog: furry, black

dog: furry, black, smooth

rabbit: furry, white, brown

dog: furry, black, brown, smooth

dog: furry, black, gray

hoop: white, red, round

dog: black, white

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

Experiments

- **Uncontrolled** Random sample of the dataset (balanced w.r.t. quantifiers)
- **Unseen Objects** Queries in the test set contain queried objects never queried in the training data.
- **Unseen Properties** Queries in the test set contain queried properties never queried in the training data.
- **Unseen O, P combination** Queries in the test set contain queried object, property combination never queried in the training data.

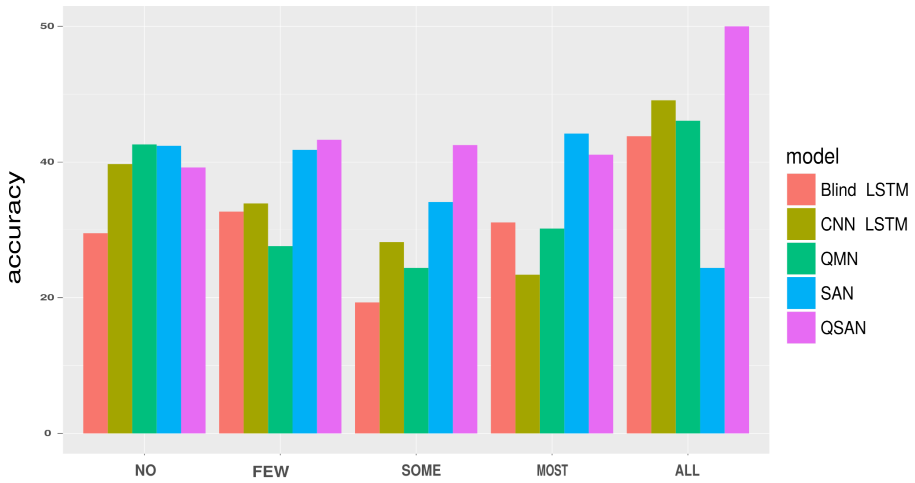
How do the models go?

Accuracies

	Q-ImageNet			
	UNC	UnsObj	UnsProp	UnsQue
Blind BOW	25.5	25.2	20.3	25.2
Blind LSTM	31.35	23.9	21.8	22.3
CNN+BOW	26.7	24.8	18.9	25.5
CNN+LSTM	34.75	23.9	20.4	22.8
SAN	37.5	26	20.5	23.4
QMN	34.1	23.2	22	**28.3**
QSAN	**45.2**	**28.6**	**22.1**	26
chance	20.0	20.0	20.0	20.0

How do the models go?

Results by quantifier



Confusion Matrix

UNC Q-ImageNet											
QSAN						**SAN**					
	no	few	some	most	all		no	few	some	most	all
no	**149**	*149*	65	7	10	no	**161**	*160*	50	9	0
few	137	**180**	69	22	8	few	*150*	**174**	61	30	1
some	54	70	**167**	65	37	some	*99*	74	**134**	83	3
most	16	23	70	**170**	*135*	most	37	65	*102*	**183**	27
all	6	11	34	*108*	**238**	all	21	40	62	*177*	97

Conjecture 1: Conclusion

Attend the restrictor than its composition with the scope

- SAN: We first showed that letting the network compose scope and restrictor on the language side, and using this representation to attend to the image, resulted in un-derperforming models.
- QMN and QSAN: Encoding into the model the fact that quantifiers express a relation between sets, to guide the attention mechanism, produced much better results.

Conjecture 1: Conclusion

Approximation is a good strategy

- precisely identifying the composition of the sets is not only beyond current state-of-the-art models but perhaps even
- detrimental to a task that is most efficiently performed by refining the *approximate numerosity* estimator of the system.
- the actual challenge of visual quantification is to find the right strategies to deal with uncertainty in object and property recognition. Humans appeal extensively to their approximate number sense to quantify.
- This may be more than an *efficiency mechanism*: as demonstrated by the QSAN models combination of soft attention and gist, approximation goes a long way in manoeuvring through the difficulties of matching words and vision.

Conjecture 1: Conclusion

Approximation is a good strategy

- precisely identifying the composition of the sets is not only beyond current state-of-the-art models but perhaps even
- detrimental to a task that is most efficiently performed by refining the *approximate numerosity* estimator of the system.
- the actual challenge of visual quantification is to find the right strategies to deal with uncertainty in object and property recognition. Humans appeal extensively to their approximate number sense to quantify.
- This may be more than an *efficiency mechanism*: as demonstrated by the QSAN models combination of soft attention and gist, approximation goes a long way in manoeuvring through the difficulties of matching words and vision.

Conjecture 1: Conclusion

Approximation is a good strategy

- precisely identifying the composition of the sets is not only beyond current state-of-the-art models but perhaps even
- detrimental to a task that is most efficiently performed by refining the *approximate numerosity* estimator of the system.
- the actual challenge of visual quantification is to find the right strategies to deal with uncertainty in object and property recognition. Humans appeal extensively to their approximate number sense to quantify.
- This may be more than an *efficiency mechanism*: as demonstrated by the QSAN models combination of soft attention and gist, approximation goes a long way in manoeuvring through the difficulties of matching words and vision.

Conjecture 1: Conclusion

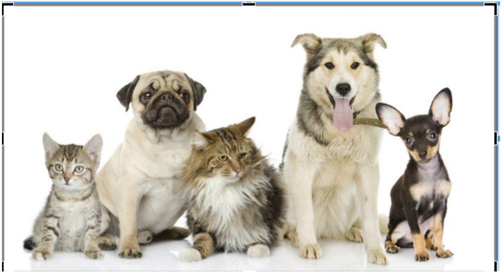
Approximation is a good strategy

- precisely identifying the composition of the sets is not only beyond current state-of-the-art models but perhaps even
- detrimental to a task that is most efficiently performed by refining the *approximate numerosity* estimator of the system.
- the actual challenge of visual quantification is to find the right strategies to deal with uncertainty in object and property recognition. Humans appeal extensively to their approximate number sense to quantify.
- This may be more than an *efficiency mechanism*: as demonstrated by the QSAN models combination of soft attention and gist, approximation goes a long way in manoeuvring through the difficulties of matching words and vision.

Layout

1. Learning quantification from images
2. Quantifiers vs. Cardinals
3. Behavioral Study

Quantifiers or Cardinals



Most of the animals are *dogs*.
vs.
Three of the animals are *dogs*.

In humans, Q vs. C underly different cognitive and neural mechanisms. What about NNs?

Dataset

Synthetic Scenarios

Pezzelle et. ali (EACL 2017) *Be Precise or Fuzzy: Learning the Meaning of Cardinals and Quantifiers from Vision*

We build a dataset of synthetic scenarios by joining together 1-9 real images from ImageNet (each image depicting one object)

- Balanced number of scenarios depicting no, few, most, all (Qs); 1,2,3,4 (Cs)
- Qs percentages defined a priori (0%, 1-49%, 51-99%, 100%, resp.)
- Train, Test differing w.r.t. different combination targets-distractors

Dataset

Combinations

Train-q				**Train-c**			
no	few	most	all	one	two	three	four
0/1	1/6	2/3	1/1	1/1	2/2	3/3	4/4
0/2	2/5	3/4	2/2	1/3	2/3	3/4	4/5
0/3	2/7	3/5	3/3	1/4	2/5	3/5	4/6
0/4	3/8	4/5	4/4	1/6	2/7	3/8	4/7
Test-q				**Test-c**			
no	few	most	all	one	two	three	four
0/5	1/7	4/6	5/5	1/2	2/4	3/7	4/8
0/8	4/9	6/8	9/9	1/7	2/9	3/9	4/9

Table: Combinations in Train and Test. targets/targets+distractors

Analysis

Only Vision: Cosine-sim(Target-Scenario) vs Dot-sim(Target-Scenario)

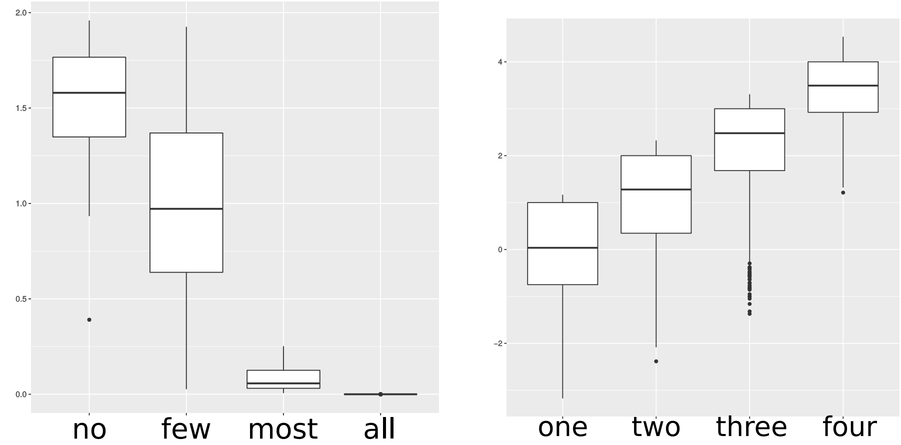
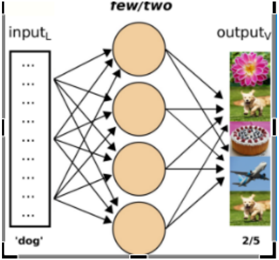


Figure: Left: quantifiers against cosine distance. Right: cardinals against dot product.

Leading idea

Q and C are (cross-modal) functions



“Few” /” Two” are matrices that given the linguistic vector of an object (dog) will retrieve the scenarios s.t. few/two of the objects are dogs.
Model Single-layer neural network (criterion: ReLU)

Q vs. C: leading idea

Learning Strategies

- **Learning strategies for Q:** it learns to obtain out of the linguistic vector of "dog" the visual vector that is most similar (based on *cosine similarity*) with the visual vectors of the scenarios with few dogs.
- **Learning strategies for C:** it learns to obtain out of the linguistic vector of "dog" the visual vector that is most similar (based on *dot product*) with the visual vectors of the scenarios with 2 dogs.
- **Intuition** Cosine is a "fuzzy" measure vs. Dot product is an "exact" measure.

Results

Cross-modal: image retrieval

	lin		**nn-cos**		**nn-dot**	
	mAP	*P2*	*mAP*	*P2*	*mAP*	*P2*
no	0.78	0.65	**0.87**	0.77	0.54	0.37
few	0.59	0.39	**0.68**	0.51	0.59	0.43
most	0.61	0.36	0.60	0.29	**0.62**	0.45
all	0.75	0.66	**1**	1	0.33	0.12
one	0.44	0.30	0.38	0.21	**0.61**	0.45
two	0.35	0.15	0.38	0.21	**0.57**	0.43
three	0.38	0.16	0.36	0.13	**0.56**	0.40
four	0.65	0.47	0.75	0.60	**0.76**	0.61

Table: R-target. *mAP* and *P2* for each model.

Conjecture 2: Conclusion

- Each Q can be represented by a multimodal function from language to vision.
- Low C can be learned by mapping language into vision.

Layout

1. Learning quantification from images
2. Quantifiers vs. Cardinals
3. Behavioral Study

On going work: What about humans?

Behavioral studies

Sandro Pezzelle, Manuela Piazza, and me

Question: which factors influence our decision to use one Q instead of another when quantity-wise they are very similar?

Currently visual factors: size of the image, color, location, cardinality, ratio.

Only-vision study:

- given a visual scene containing animals and artifacts,
- subjects have to choose the Q out of 9 options: none, almost none, very few, few, some, many, most, almost all, all

On going work: What about humans?

Behavioral studies

Sandro Pezzelle, Manuela Piazza, and me

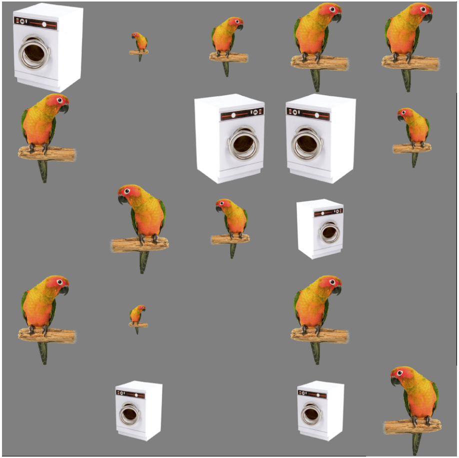
Question: which factors influence our decision to use one Q instead of another when quantity-wise they are very similar?

Currently visual factors: size of the image, color, location, cardinality, ratio.

Only-vision study:

- given a visual scene containing animals and artifacts,
- subjects have to choose the Q out of 9 options: none, almost none, very few, few, some, many, most, almost all, all

Example



Conclusion

- **Conjecture 1:** we can learn their *litteral meaning* (respecting the abstract scale) from *images*.
- Yes, by creating the gists of the compared sets.
- **Conjecture 2:** they can be represented by a *cross-modal function*.
- Yes, from the word embedding of the noun to the visual scene, using cosine as objective.
- **Conjecture 3:** text corpora could help learn their *use*.
- Still unexplored

Conclusion

- **Conjecture 1:** we can learn their *litteral meaning* (respecting the abstract scale) from *images*.
- Yes, by creating the gists of the compared sets.
- **Conjecture 2:** they can be represented by a *cross-modal function*.
- Yes, from the word embedding of the noun to the visual scene, using cosine as objective.
- **Conjecture 3:** text corpora could help learn their *use*.
- Still unexplored

Conclusion

- **Conjecture 1:** we can learn their *litteral meaning* (respecting the abstract scale) from *images*.
- Yes, by creating the gists of the compared sets.
- **Conjecture 2:** they can be represented by a *cross-modal function*.
- Yes, from the word embedding of the noun to the visual scene, using cosine as objective.
- **Conjecture 3:** text corpora could help learn their *use*.
- Still unexplored

The team

Ionut

Sandro

Aurelie

me

Descriptive statistics of the two Q datasets

	Q-COCO	**Q-ImageNet**
unique objects	29	161
unique properties	44	24
properties per object (mean)	15.7	8.0
objects per property (mean)	10.34	53.67
objects per scenario (mean)	8.49	16
objects per scenario (min-max)	6 - 22	16 - 16
BBs per object (mean)	826.14	48.38
BBs per object (min-max)	16 - 4741	13 - 1149
BBs per property (mean)	2,090.39	728.12
BBs per property (min-max)	616 - 8,320	23 - 2,689
total images	2,888	7,790
total BBs	23,958	7,790
total queries	58,673	40,000

Table: Descriptive statistics for Q-COCO and Q-ImageNet datasets.

Vector Representations

Visual input For each bounding box in each scenario, we extract a visual representation using a Convolutional Neural Network. We use the VGG-19 model pre-trained on the ImageNet ILSVRC data and the MatConvNet toolbox for features extraction. Each bounding box is represented by a 4096-dimension vector extracted from the 7th fully connected layer (fc7). For computational efficiency, we subsequently reduce the vectors to 400 dimensions by applying Singular Value Decomposition (SVD).
Linguistic input Similarly, each word in a query is represented by a 400-dimension vector built with the Word2Vec CBOW architecture, using the parameters that were shown to perform best in Baroni et. al 2014. The corpus used for building the semantic space is a 2.8 billion tokens concatenation of the web-based UKWaC, a mid-2009 dump of the English Wikipedia, and the British National Corpus (BNC).

Results by ratios

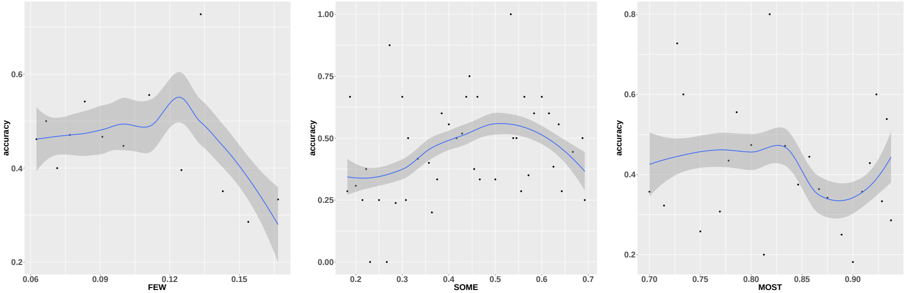


Figure: QSAN. Accuracy in UNC plotted against the ratios of target objects over restrictors. Left: ‘few’. Center: ‘some’. Right: ‘most’.

CNN+BOW

This model is an adaptation of iBOWIMG.
It uses the same linguistic input as BOW above, concatenated with a visual input. As in BOW, the query question is first converted to a one-hot bag-of-words vector, which is further transformed into a 'word feature' embedding.
This linguistic embedding is concatenated with an 'image feature' obtained from a convolutional neural network (CNN). The resulting embedding is sent to a softmax classifier which predicts one of five quantifiers, as above.
In order to have one single vector for the visual input, we simply concatenate the visual vectors of the individual bounding boxes in each one of our scenarios. For the Q-COCO dataset, where the number of objects contained in one images ranges from 6 to 22, we concatenate our 'frozen' visual vectors into a 8,800-dimension vector (i.e. 22*400 dimensions) and we fill the 'empty' cells of the scenario with zero vectors.