Imparare a quantificare guardando
Learning to quantify by watching

Sandro Pezzelle[1], Ionut Sorodoc[2], Aurelie Herbelot[1], and Raffaella Bernardi[1]

[1]University of Trento
[2]EMLCT

[1]*{firstname.lastname}@unitn.it*
[2]*ionut.sorodoc@gmail.com*

CLiC-IT, Naples
December 5th 2016

Outline

1. Overview
2. Data
3. Models
4. Experiment
5. Conclusions

Abstract

- Multimodal model quantifying over visual scenes using natural language **quantifiers** (*no*, *few*, *some*, *most*, *all*)

Abstract

- Multimodal model quantifying over visual scenes using natural language **quantifiers** (*no*, *few*, *some*, *most*, *all*)
- Visual Question Answering (**VQA**) task with genuine understanding of both linguistic and visual inputs

Task

Task

How many **dogs** are **black**? No/few/some/most/all?

Dataset

What is needed

Visual scenes containing multiple objects w/ various properties

Dataset

What is needed

Visual scenes containing multiple objects w/ various properties

- From ImageNet, pics labeled wrt object (*dog*) and properties (*black*)

Dataset

What is needed

Visual scenes containing multiple objects w/ various properties

- From ImageNet, pics labeled wrt object (*dog*) and properties (*black*)
- Filtering based on N properties, frequency of corresponding word

Dataset

What is needed

Visual scenes containing multiple objects w/ various properties

- From ImageNet, pics labeled wrt object (*dog*) and properties (*black*)
- Filtering based on N properties, frequency of corresponding word
- Selected 161 different objects (7324 images, 24 properties)

Dataset

What is needed

Visual scenes containing multiple objects w/ various properties

- From ImageNet, pics labeled wrt object (*dog*) and properties (*black*)
- Filtering based on N properties, frequency of corresponding word
- Selected 161 different objects (7324 images, 24 properties)
- Built synthetic (plausible) scenarios made up of 16 different images

Dataset

What is needed

Visual scenes containing multiple objects w/ various properties

- From ImageNet, pics labeled wrt object (*dog*) and properties (*black*)
- Filtering based on N properties, frequency of corresponding word
- Selected 161 different objects (7324 images, 24 properties)
- Built synthetic (plausible) scenarios made up of 16 different images
- Built datapoints: <scenario, query, answer>

Materials

Visual features

4096-d features extracted from *fc7* of **CNN** (VGG-19 pretrained on Imagenet)

Materials

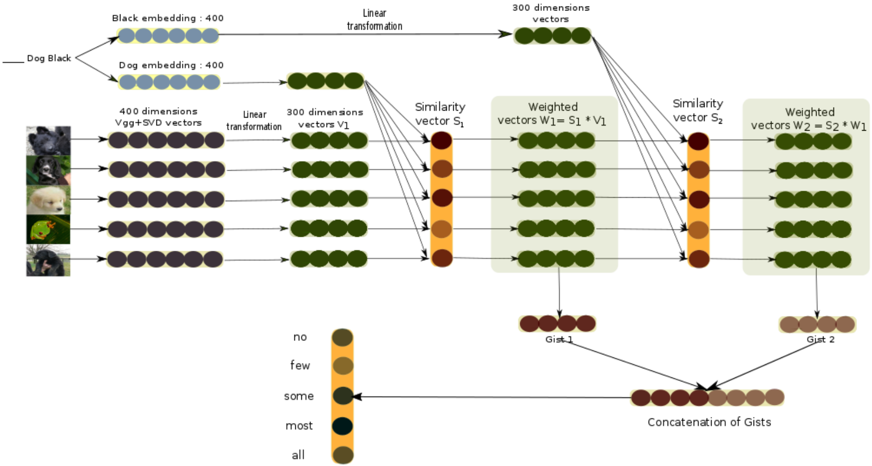
Visual features

4096-d features extracted from *fc7* of **CNN** (VGG-19 pretrained on Imagenet)

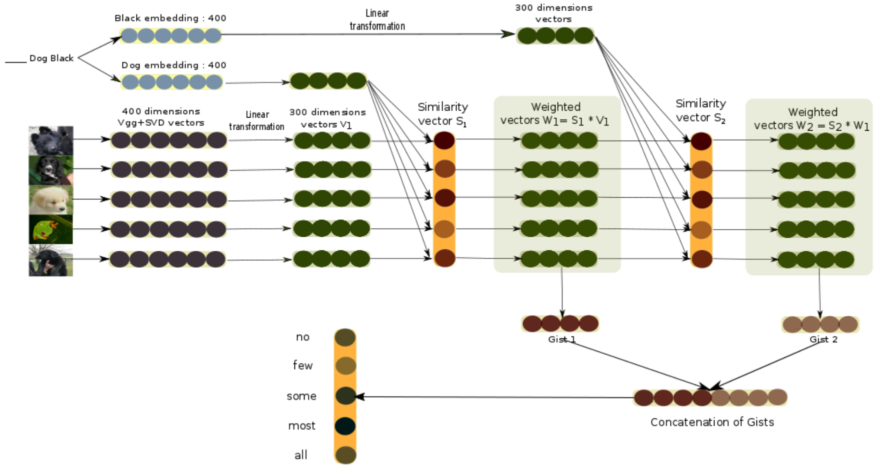
Word embeddings

400-d **word2vec** embeddings built with CBOW on 2.8B token corpus

Quantifier Memory Network (qMN) model



Quantifier Memory Network (qMN) model



Baseline

VQA state-of-art `iBOWIMG` (Zhou et al., 2015)

Experimental settings

Uncontrolled

10,000 datapoints randomly split in train (70%), val (10%), and test (20%)

Experimental settings

Uncontrolled

10,000 datapoints randomly split in train (70%), val (10%), and test (20%)

Unseen queries

7,000 datapoints selected for train, val and test w/ same scenarios and objects but unseen properties

Experimental settings

Uncontrolled

10,000 datapoints randomly split in train (70%), val (10%), and test (20%)

Unseen queries

7,000 datapoints selected for train, val and test w/ same scenarios and objects but unseen properties

Unseen scenarios

7,000 datapoints selected for train, val and test w/ same objects and properties but unseen scenarios

Results

	Unseen queries		Unseen scenarios		Uncontrolled	
	qMN	iBOWIMG	qMN	iBOWIMG	qMN	iBOWIMG
some	**43.08**	25.8	32.62	**39.83**	18.16	**22.13**
all	**67.06**	61.42	**50.51**	34.1	**52.22**	40.34
no	77.5	**96.52**	**67.99**	50.33	**59.7**	49.5
few	**38.01**	23.96	25.86	**26.84**	**32.25**	21.25
most	**46.97**	25.27	**39.25**	29.17	**32.14**	20.4

Table: Percentage of target quantifiers correctly predicted by each model

Error analysis

qMN					
	some	all	no	few	most
some	73	<u>88</u>	57	<u>89</u>	<u>95</u>
all	29	**211**	20	19	<u>125</u>
no	32	28	**240**	70	32
few	46	53	<u>104</u>	**129**	68
most	49	<u>148</u>	31	38	126
iBOWIMG					
	some	all	no	few	most
some	89	77	50	<u>108</u>	78
all	45	**163**	63	46	<u>87</u>
no	30	69	**199**	59	52
few	<u>82</u>	<u>81</u>	<u>100</u>	<u>85</u>	52
most	75	<u>110</u>	63	64	80

Table: Confusion matrices for `qMN` and `iBOWIMG`

Qualitative analysis

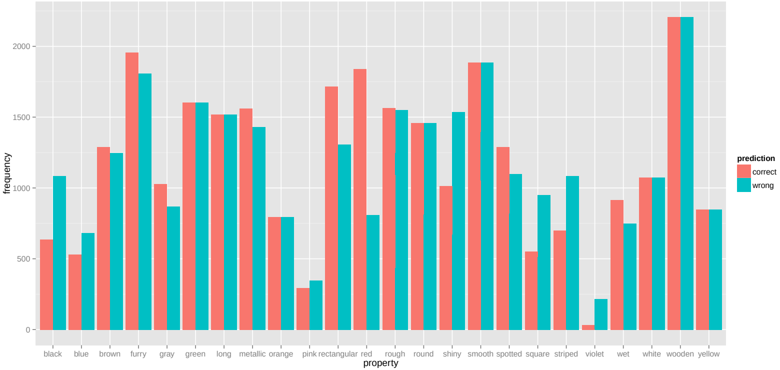


Figure: Correct/wrong cases wrt frequency of noun-property pair (Unc setting)

Discussion

- Our qMN model significantly outperforms the baseline in all three settings (around 8% better)

Discussion

- Our qMN model significantly outperforms the baseline in all three settings (around 8% better)
- Quantification cannot be handled by simply memorizing correlations (iBOWIMG fails)

Discussion

- Our qMN model significantly outperforms the baseline in all three settings (around 8% better)
- Quantification cannot be handled by simply memorizing correlations (iBOWIMG fails)
- Proper understanding of both visual and linguistic input and their interaction is needed

Discussion

- Our qMN model significantly outperforms the baseline in all three settings (around 8% better)
- Quantification cannot be handled by simply memorizing correlations (iBOWIMG fails)
- Proper understanding of both visual and linguistic input and their interaction is needed
- “Logical” quantifiers (*no*, *all*) are easier to learn than “proportional” ones (*most* and *few*).

Future research

- Experiment with more natural datasets (i.e. real scenes)

Future research

- Experiment with more natural datasets (i.e. real scenes)
- Collect human judgments on quantifiers' *use* to take into account pragmatics beyond "proportions"

Future research

- Experiment with more natural datasets (i.e. real scenes)
- Collect human judgments on quantifiers' *use* to take into account pragmatics beyond "proportions"
- Test "fuzzy" against "precise" quantification (quantifiers vs. exact cardinals)

Thank you!

("all" the authors)