A Survey of Current Datasets for Vision & Language Research

Francis Ferraro[1], Nasrin Mostafazadeh[2], Ting-Hao (Kenneth) Huang[3], Lucy Vanderwende, Jacob Devlin, Michel Galley, Margaret Mitchell

1 Johns Hopkins University, 2 University of Rochester, 3 Carnegie Mellon University

Microsoft Research

Motivation

Recent explosion in vision&language work, from captioning and video description to question answering and beyond.

What is lacking? A systematic way to objectively analyze the quality of the datasets:

- **Language Quality Criteria**
- Vision Quality Criteria

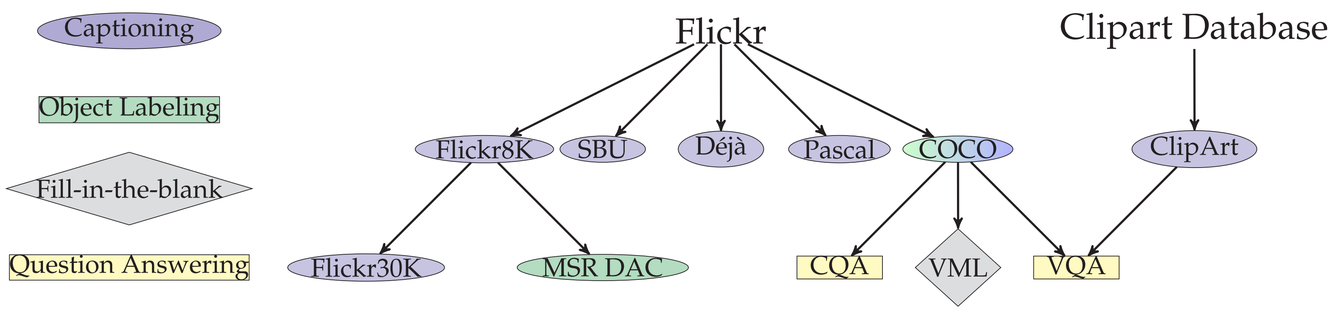
VisionAndLanguage.net

Our evolving website, `http://visionandlanguage.net`, contains pointers and references to a extensive list of older and newly expanding vision &language datasets:

- Categorizes evolving research tasks introduced in vision & language.
- Makes record of the major available datasets and their characteristics.

Source on GitHub, allowing for community involvement in updating.

Categorization of Recent V&L Datasets



Flickr8K [1]; Flickr30K [2]; SBU [3]; MSR **D**ense **A**nnotation **C**orpus [4]; Déjà Images [5]; Pascal [6]; Microsoft **C**ommon **O**bjects in **Co**ntext [7]; Toronto **C**OCO-**Q**uestion **A**nswering [8]; **V**isual **M**ad**L**ibs [9]; **V**isual **Q**uestion **A**nswering [10]; Abstract Scenes **ClipArt** [11]

Reporting Bias

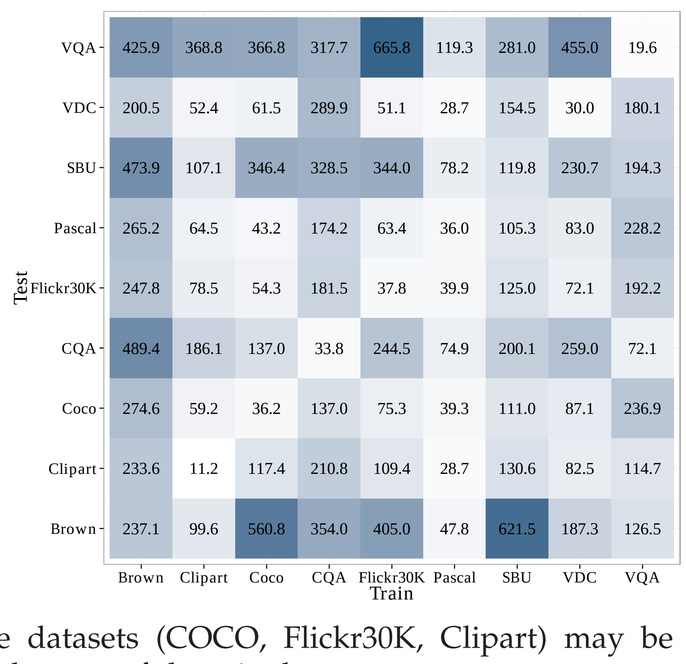
Reporting bias and **Photographer's bias**: People are selective of what they "report" (Gordon&Van Durme, 2013; Torralba&Efros, 2011).

Case Study: Dense Annotation Corpus (Yatskar et al., 2014)

- 8.04 visible objects per image.
- Only 2.7 (34%) are appear in a caption.

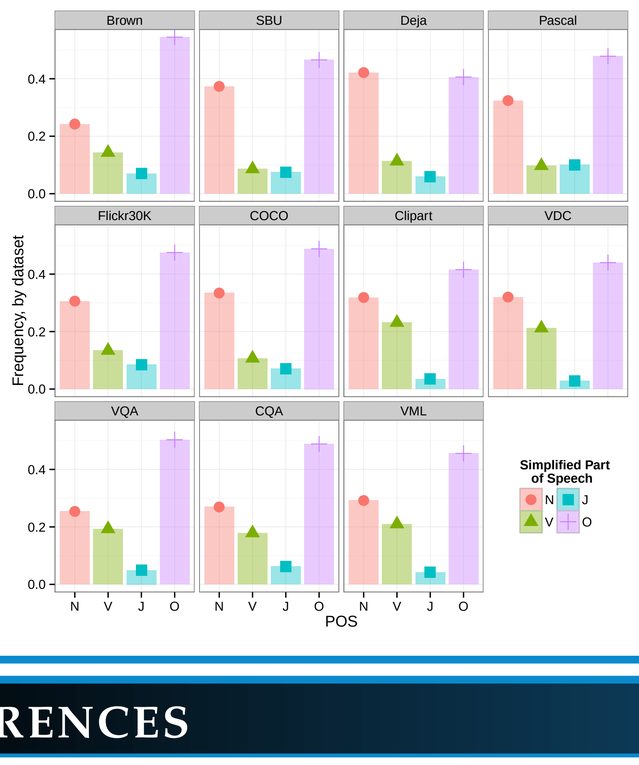
Many selection biases present in abstract scenes are also present in photos.

Fine-grained Perplexity



Some datasets (COCO, Flickr30K, Clipart) may be more useful as out-of-domain data.

POS Distribution



Language Quality

The distinction between a qualitatively "good" or "bad" dataset is task dependent; all criteria should be viewed through the lens of particular downstream tasks.

- **Vocabulary Size**
- **Average Sentence Length**
- **Part of Speech Distribution**
- **Syntactic Complexity** (Yngve, Frazier): measures embedding/branching in a sentence's syntax.
- **Abstract/Concrete Ratio**: Percentage of abstract terms (from Vanderwende et al. (2015)) indicates the range of visual and non-visual concepts the dataset covers.
- **Perplexity**: Compared against 5-gram language model learned on generic 30B words English.

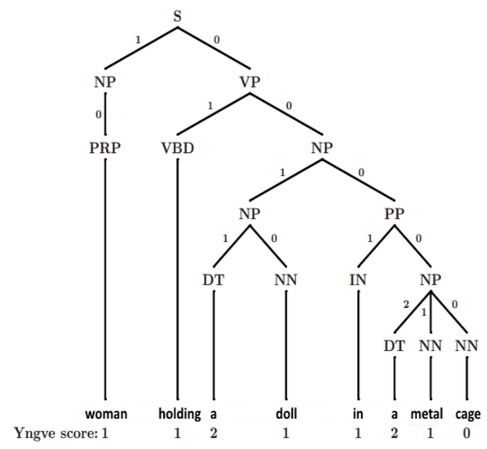


Figure: Example parse tree with branch scores for Yngve scoring.

		Size(k)		Language					
	Dataset	Img	Txt	Frazier	Yngve	Vocab Size (k)	Sent Len.	%Abs	Ppl
Balanced	Brown	-	52	18.5	77.21	47.7	20.82	15.24%	194
User-Gen	SBU	1000	1000	9.70	26.03	254.6	13.29	3.74%	346
	Deja	4000	180	4.13	4.71	38.3	4.10	9.70%	184
Crowd-sourced	Pascal	1	5	8.03	25.78	3.4	10.78	17.74%	123
	Flickr30K	32	159	9.50	27.00	20.3	12.98	14.98%	118
	COCO	328	2500	9.11	24.92	24.9	11.30	12.96%	121
	Clipart	10	60	6.50	12.24	2.7	7.18	17.96%	126
Video	VDC	2	85	6.71	15.18	13.6	7.97	12.86%	148
Beyond	VQA	10	330	6.50	14.00	6.2	7.58	19.22%	113
	CQA	123	118	9.69	11.18	10.2	8.65	16.14%	199
	VML	11	360	6.83	12.72	11.2	7.56	17.19%	110

Vision Quality Criteria in the Paper!

References

1. Rashtchian et al., (2010)
2. Young et al., (2014)
3. Ordonez et al., (2011)
4. Yatskar et al., (2014)
5. Chen et al., (2015)
6. Farhadi et al., (2010)
7. Lin et al., (2014)
8. Ren et al., (2015)
9. Yu et al., (2015)
10. Antol et al., (2015)
11. Zitnick et al., (2013)