Compositional Generalisation in Image Captioning

Desmond Elliott
University of Copenhagen

Mitja Nikolaus, Mostafa Abdou, Rahul Aralikatte, Matthew Lamm, Desmond Elliott. CoNLL 2019

CLASP/CLT Seminar
February 20, 2020

Image captioning[1]

Model

An big white dog is sitting on a bench just like a human

[1]See Bernardi et al (JAIR 2016) for a more comprehensive overview

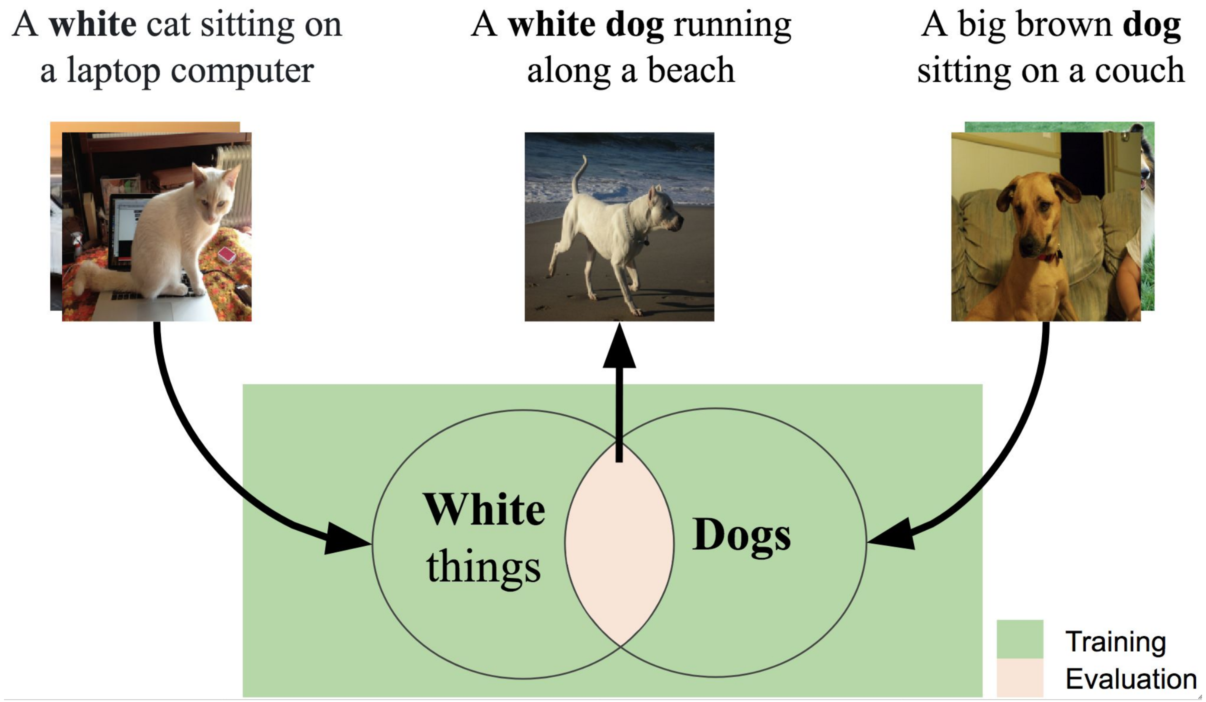
On Surpassing Human-level Performance

- State-of-the-art models **surpass human performance** according to BLEU, etc.[2]
- But **humans still prefer** the human-written sentences (e.g. Vinyals et al., 2017)
- Systematic compositionality is a key property of human language (Partee, 1984)

How well do image captioning models perform X+Y generalisation, e.g. *adjective-noun* and *noun-verb*?

[2] http://cocodataset.org/#captions-leaderboard

Forcing Paradigmatic Gaps



Related Work: Grounded Language

- `<subject, relation, object>` prediction (Atzmon et al. 2016)
- CLEVR: synthetic data question-answering (W?-words) (Johnson et al. 2016)
- Quantification in grounding, e.g. some, all, few, many... (Pezzelle et al. 2018)
- Captioning unseen objects (Agrawal et al. 2019), unseen object pairs (Lu et al. 2018)

This work: natural images and full sentence generation

Generalization of 24 Concept Pairs

- Choose concepts that are *likely* to be encoded by a visual recognition model

black cat	big bird	red bus	small plane
eat man	lie woman	white truck	small cat
brown dog	big plane	ride woman	fly bird
white horse	big cat	blue bus	small table
hold child	stand bird	black bird	small dog
white boat	stand child	big truck	eat horse

Adjectives: Colour Size

Verbs: Transitive Intransitive

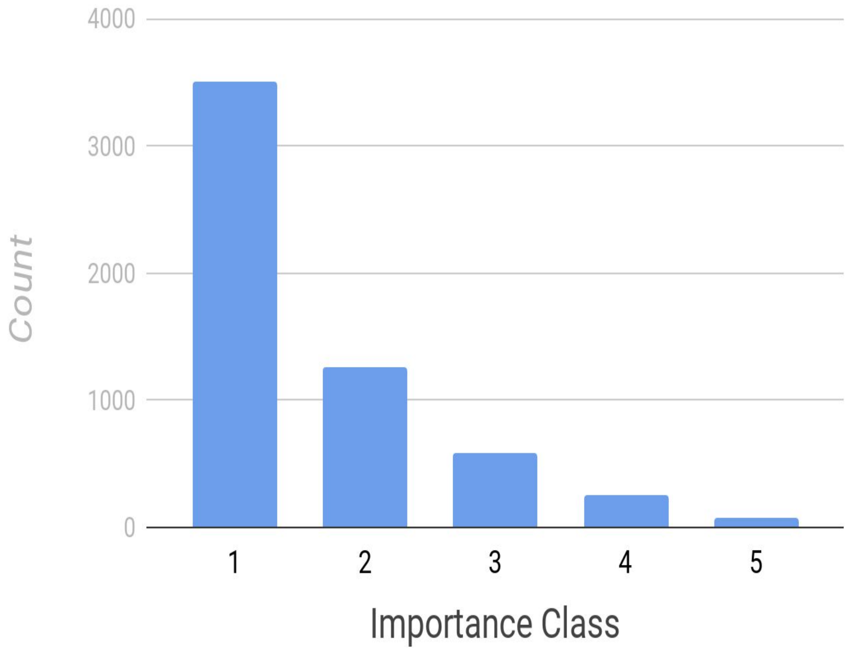
Nouns: Animate Inanimate

Note: These are only intersective adjectives (Partee, 1995)

Importance Classes of Concept Pairs

- The distribution of the concept pairs shows their **importance class** (van Miltenburg et al. 2018)

- Skewed distribution is likely because the caption crowdsourcing task is relatively loosely defined



Methodology

- English Dataset: MS COCO
  - **Train** on data that excludes the held-out concept pairs
  - **Evaluate** on data that contains only the held-out concept pairs
- State-of-the-art models:
  - Show, Attend and Tell (SAT; Xu et al., 2015)
  - Bottom-Up and Top-Down (BUTD; Anderson et. al., 2018)
  - Image-Sentence Ranking: (VSE++; Faghri et al., 2018)
- **Full Data**: Train on the full MS COCO dataset

Recall@K Evaluation Metric

a big **white dog** is on a bench
a man is walking a dog in the street
a man is walking a **white dog**
a man is riding a white horse
a boy is sitting next to a **white dog**

Evaluation metric: Recall of the **concept pair** in the *top K* generated captions

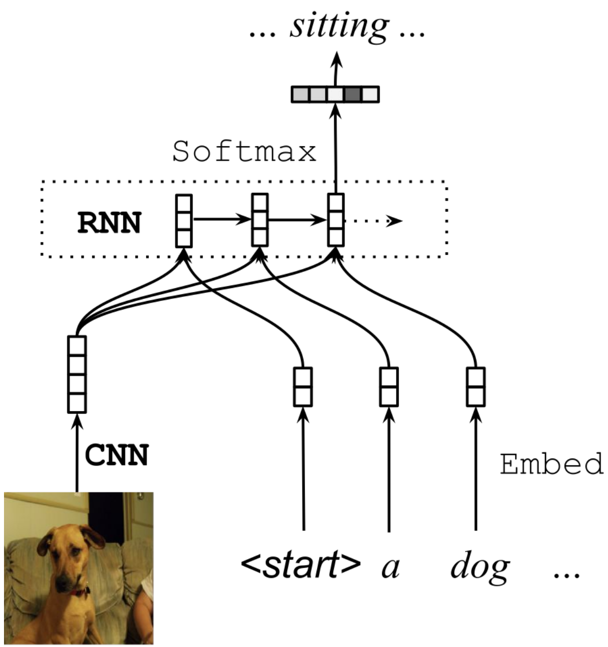
K x M generated captions

set of generated captions that contain the concept pair

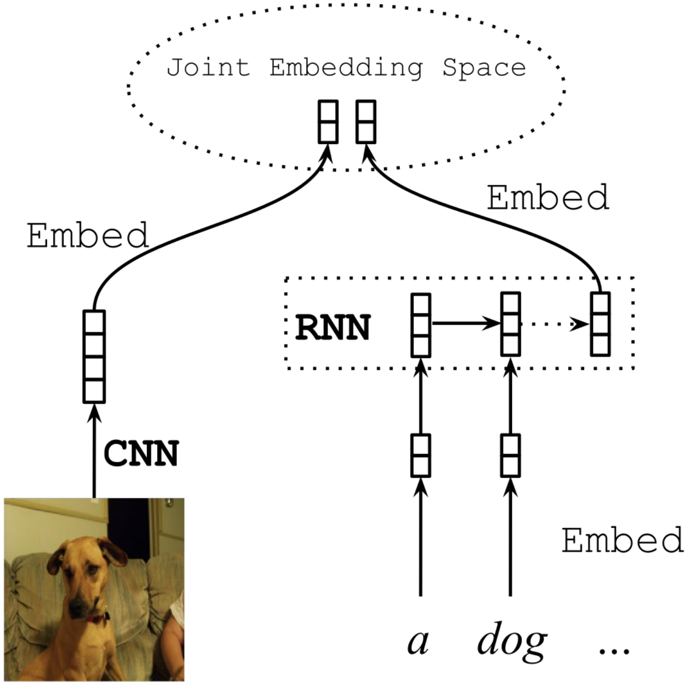
$$\text{Recall@K} = \frac{\left|\left\{\langle s_k^m \rangle \mid \exists k{:} s_k^m \in \mathcal{C}\right\}\right|}{M}$$

number of images in the evaluation set

Captioning



Ranking



Evaluating the Ranking Model

enc(`A huge` **`white dog`** `is sitting on a bench.`)

enc()

Similarity

`+ 500 images of white things`

`+ 500 images of dogs`

State of the Art Performance

Caption Generation

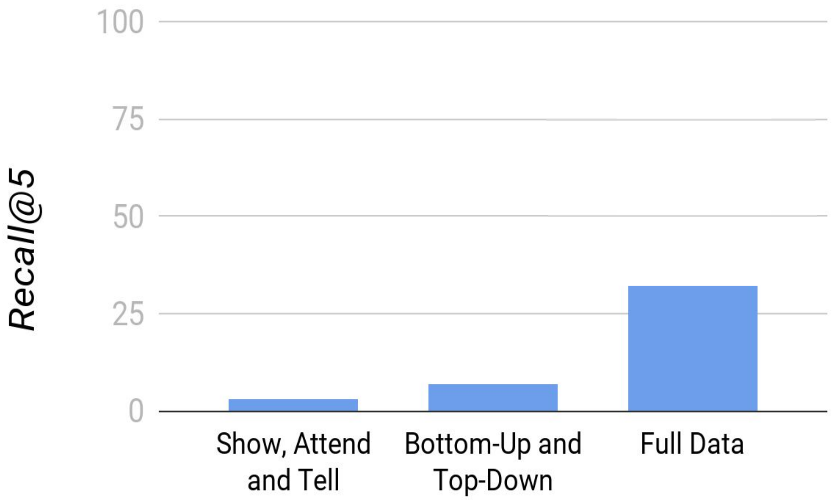
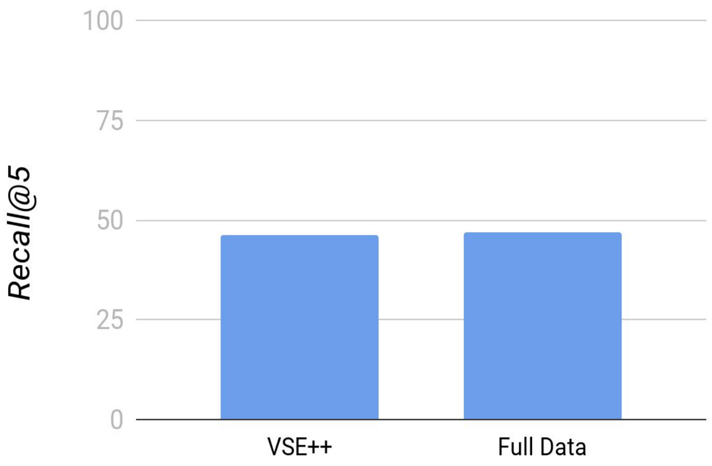


Image-Sentence Ranking

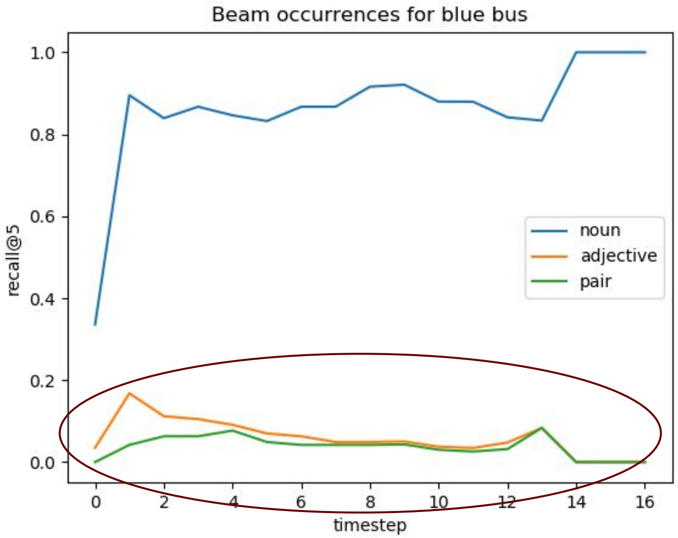


Analysis: rainbow dog

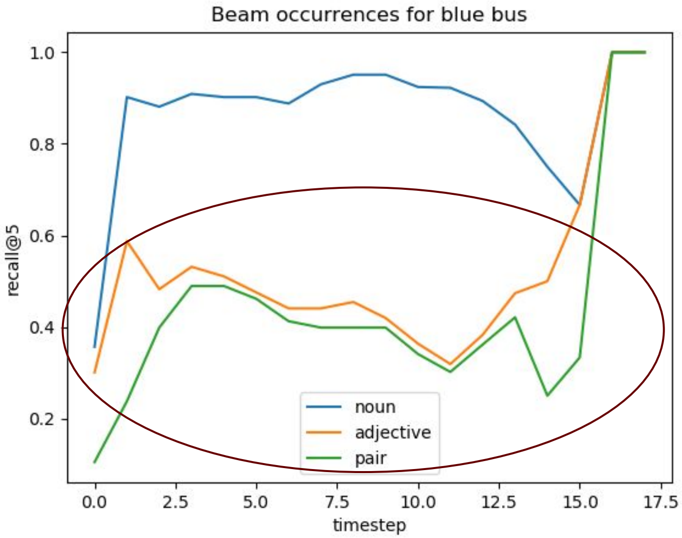
	log(prob)
a small dog sitting in a car seat	-7.57
a small **brown** dog sitting in a car seat	-11.38
a small **black** dog sitting in a car seat	-10.23
a small **white** dog sitting in a car seat	-12.37
a small **gray** dog sitting in a car seat	-12.16
a small **red** dog sitting in a car seat	-14.15
a small **blue** dog sitting in a car seat	-14.34
a small **yellow** dog sitting in a car seat	-15.98
a small **green** dog sitting in a car seat	-16.83

More analysis: peeking at the caption beam

Bottom-Up and Top-Down



Full Data

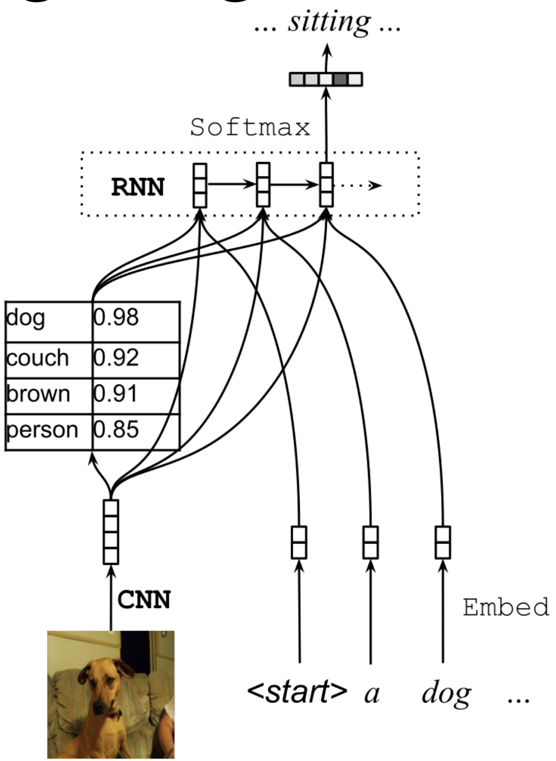


Even More Analysis: Probing Image Attributes

Semantic Compositional Network[3]

- Average attribute probabilities:
  - `dog`: 0.98
  - `brown`: 0.74
- Recall@5: 0

→ Problem is **not the encoder**



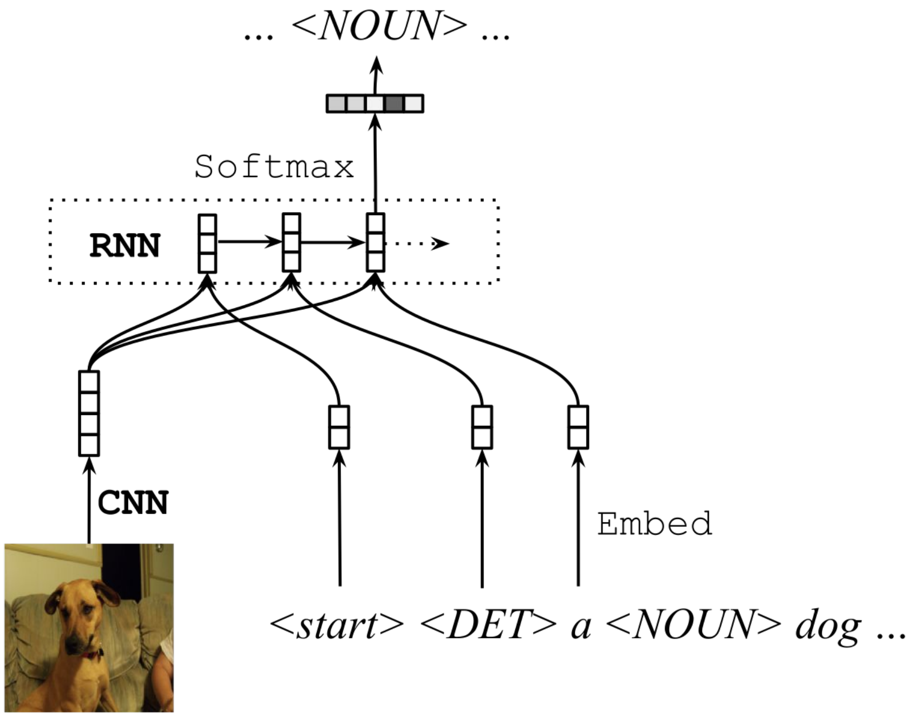
[3] Gan et. al. (2017)

Bonus Analysis: Forcing “Structure”

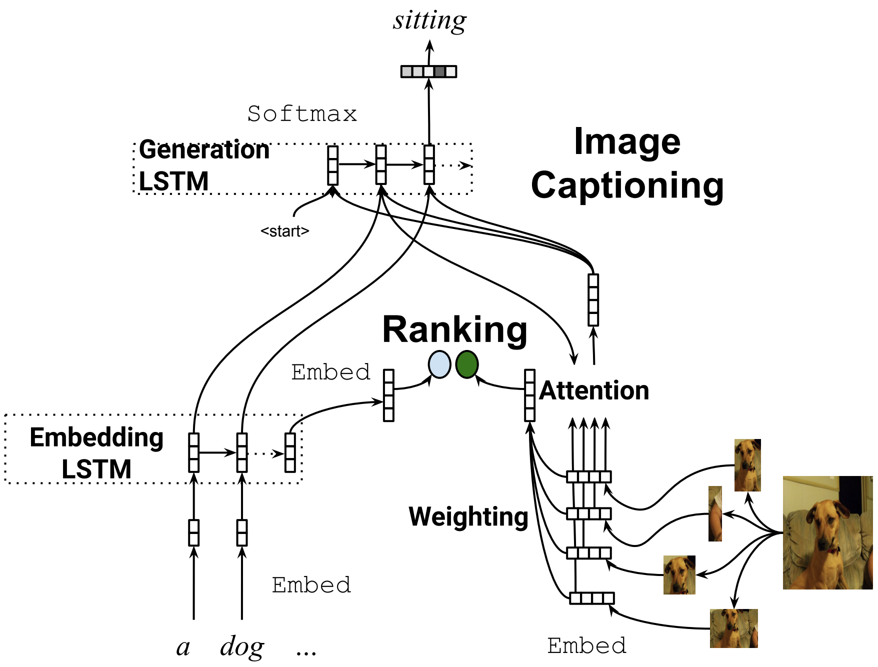
Adapted dataset: Interleave POS tags

- Force an adjective *before* noun
  - Recall@5 of up-to 17.8

→ Problem is the **decoding strategy**



BUTR: Multi-task Captioning and Image-Sentence Ranking



$$\boldsymbol{h}_t^l = \mathrm{LSTM}(\boldsymbol{W}_1 \boldsymbol{o}_t, \boldsymbol{h}_{t-1}^l)$$

Encode tokens

$$\boldsymbol{s}^* = \boldsymbol{W}_2 \boldsymbol{h}_{t=T}^l$$

Sentence: embedded

d image

$$\boldsymbol{v}_r^e = \boldsymbol{W}_3 \boldsymbol{v}_r$$
$$\beta_r' = \boldsymbol{W}_4 \boldsymbol{v}_r^e$$
$$\boldsymbol{\beta} = \mathrm{softmax}(\boldsymbol{\beta}')$$

Embed each image region

$$\boldsymbol{v}^* = \sum_{r=1}^{R} \beta_r \boldsymbol{v}_r^e$$

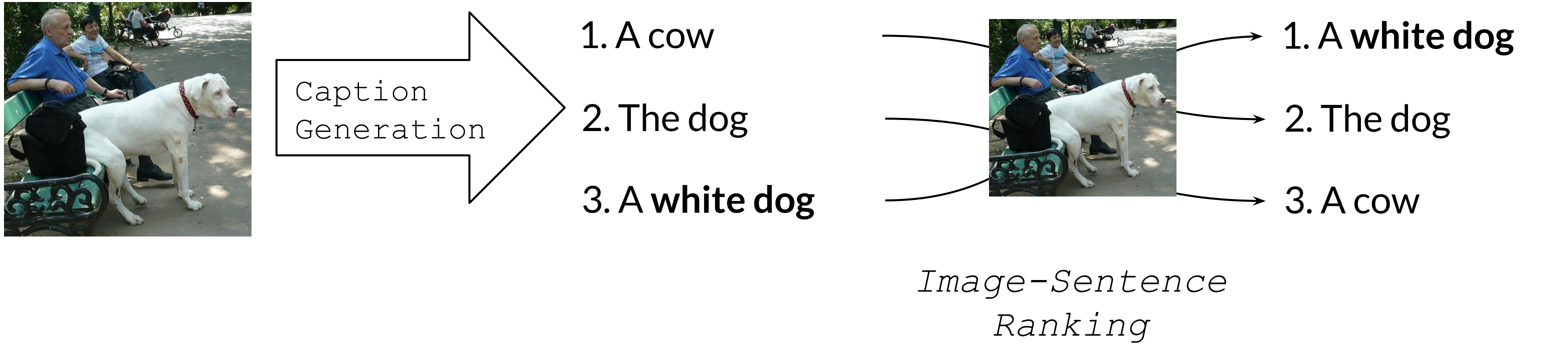
Image: Unconditional weighted sum of embeddings

$$\mathcal{L}_{\mathrm{rank}}(\theta_2) = \max_{s'}[\alpha + \cos(i, s') - \cos(i, s)]_+ + \max_{i'}[\alpha + \cos(i', s) - \cos(i, s)]_+$$

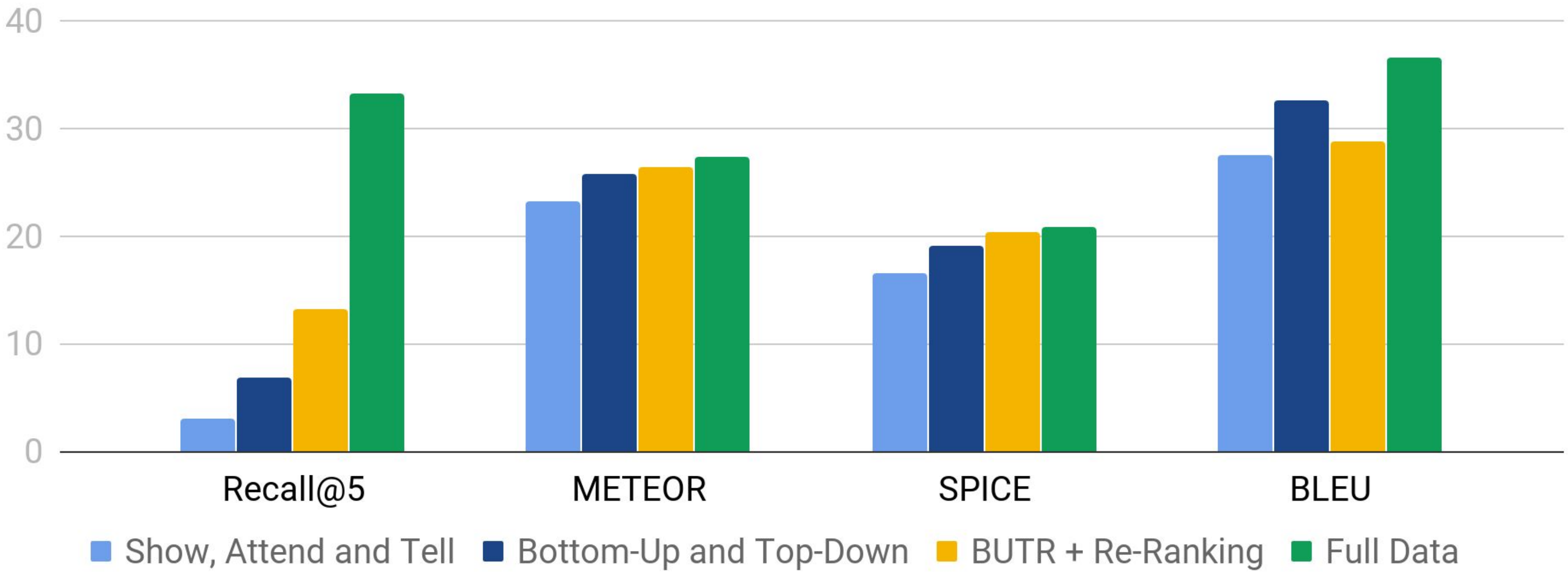
VSE++ ranking model (Faghri et al, BMVC18); BUTD captioning model (Anderson et al. CVPR18); GradNorm (Chen et al. ICML18)

BUTR: Reranking the beam

1. Generate *K* candidate captions
2. Re-rank using image-sentence similarity ranking model



Results



40
30
20
10
0
Recall@5
METEOR
SPICE
BLEU
Show, Attend and Tell
Bottom-Up and Top-Down
BUTR + Re-Ranking
Full Data

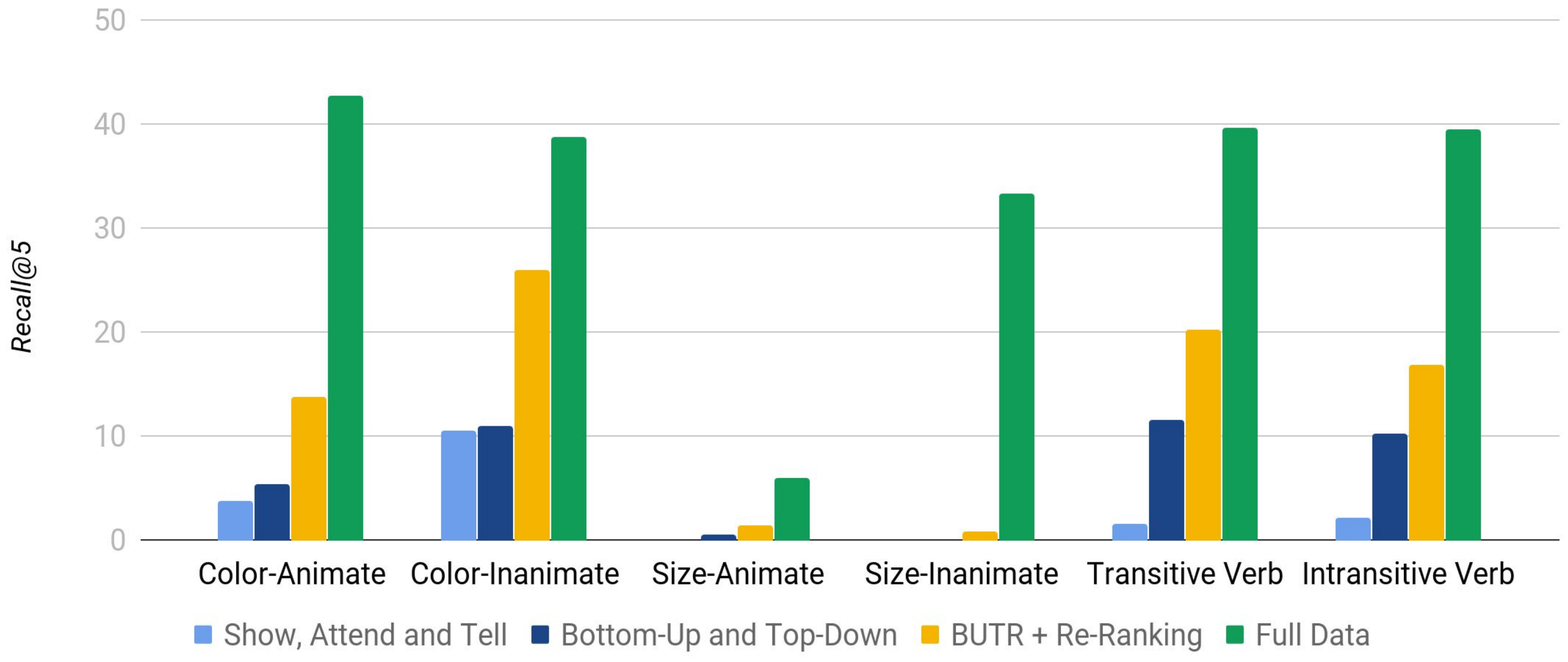
BLEU blues: output diversity analysis[3]

- Generating novel sentences is key to success in this task

Model	ASL	SDSL	Types	TTR_1	%Novel
Liu et al. (2017)	**10.3**	1.32	598	0.17	50.1
Vinyals et al. (2017)	10.1	1.28	953	0.21	90.5
Shetty et al. (2017)	9.4	1.31	**2611**	0.24	80.5
BUTD	9.0	1.01	1162	0.22	56.4
BUTR	10.2	**1.76**	1882	**0.26**	**93.6**
Validation data	11.3	2.61	9200	0.32	95.3

[3]We used the toolkit from van Miltenburg et al. (COLING 2018)

Recall@5 by Type

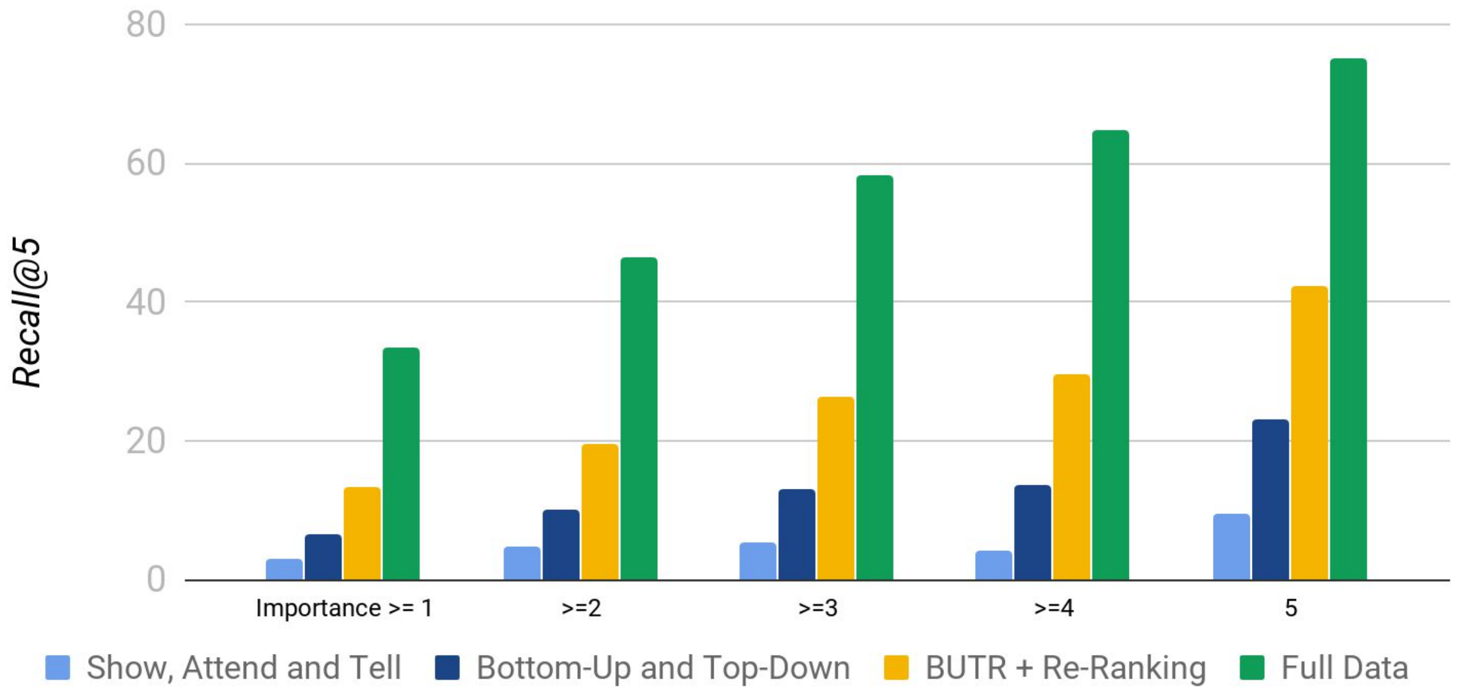


How big is a big thing anyway?

- Size modifiers are more concerned with reference classes than the depicted size.

Concept	Average bounding box size (in pixels)
small cat	42,920.6 ± 38,952.2
big cat	44,057.4 ± 41,979.9
small plane	33,718.8 ± 30,481.2
big plane	33,263.1 ± 31,722.9
small dog	36,939.5 ± 41,073.3
big dog	37,098.3 ± 40,088.6
small table	80,762.0 ± 89,751.0
big table	72,958.0 ± 91,340.0
small bird	15,063.0 ± 19,487.6
big bird	14,707.8 ± 27,008.7
small truck	30,014.0 ± 49,121.4
big truck	32,918.2 ± 46,379.8

Performance by Importance Class



Qualitative Analysis (beam k=1)

	white horse	small plane	bird stand
			
SAT	a black and white cow standing on top of a lush green field	a fighter jet on top of a lush green field	a white bird sitting on top of a car
BUTD	a brown and white cow standing on a lush green field	a white and green airplane on a field	a white bird sitting on top of a car
BUTR+ Re-Ranking	a large **white horse** standing on top of a green field	a white and green plane is parked on the grass	a large white **bird standing** on top of a car

Conclusions

- State-of-the-art image captioning models *do not* compositionally generalise

- Jointly learned ranking improves generalisation *and* text-similarity measures
    - Solution is **not** specific to generalisation of pairs of concepts

- Future work
    - Tackle generalisation with syntactic planning
    - Improve size modifier generalization
    - Integrate jointly-trained discriminative re-rankers in other NLP tasks