Towards image captioning and evaluation

Vikash Sehwag, Qasim Nadeem

Overview

- Why automatic image captioning?
- Overview of two caption evaluation metrics
- **Paper:** Captioning with Nearest-Neighbor approaches
  - Motivation
  - Description
  - Evaluation
- **Paper:** SPICE- a new caption evaluation metric
  - Motivation
  - Methods
  - Evaluation

Why do need Captioning?

- Image captioning for the **visually impaired**

Automated Neural Image Caption Generator for Visually Impaired People

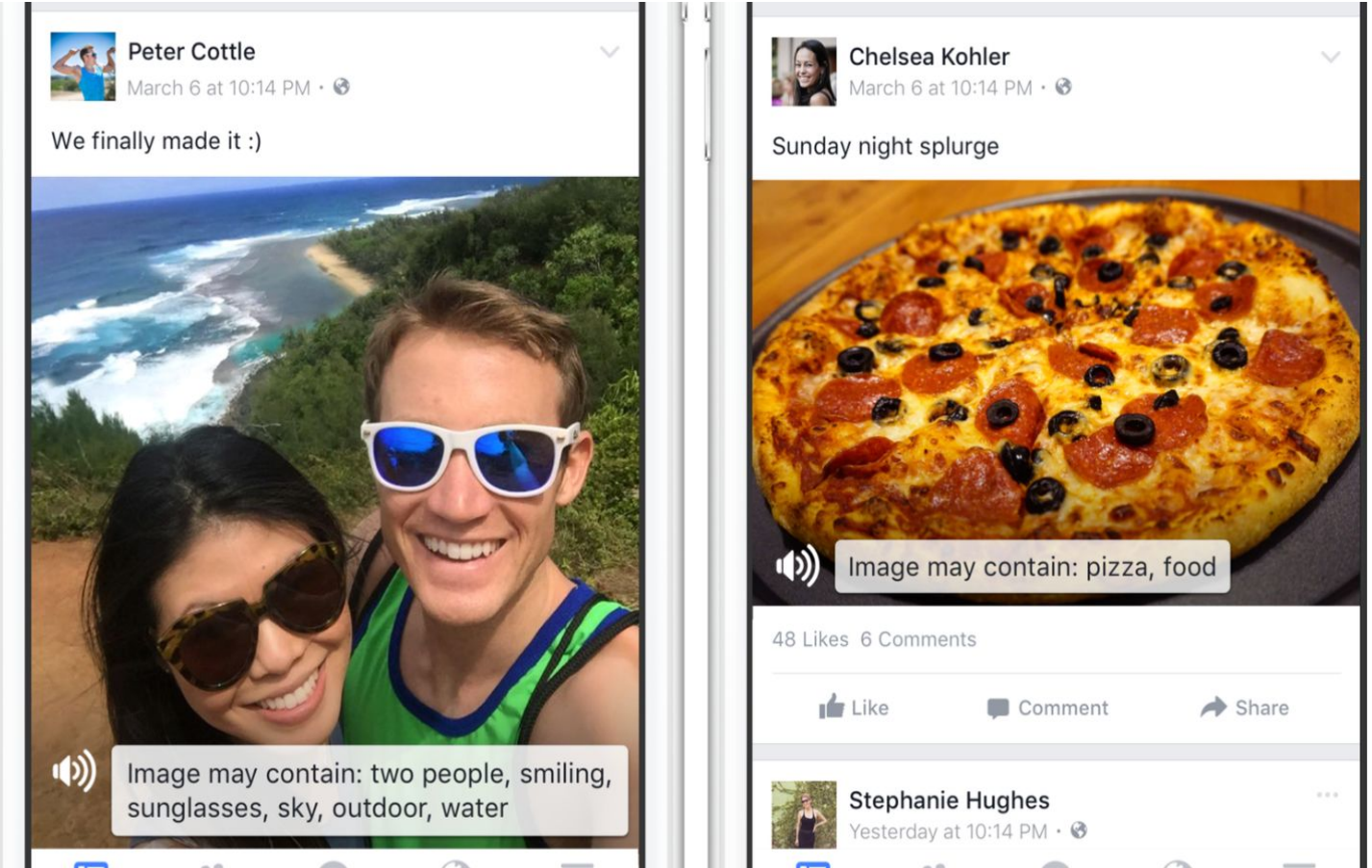
---

Christopher Elamri, Teun de Planque
Department of Computer Science
Stanford University

- Fed **penultimate layer** of a CNN into a vanilla RNN or an LST to generate valid english captions **(2016)**

Why do need Captioning?

- Facebook's **Alternative-text** can be improved



Credits:

https://www.theverge.com/2016/4/5/11364914/facebook-automatic-alt-tags-blind-visually-impared

https://code.facebook.com/posts/457605107772545/under-the-hood-building-accessibility-tools-for-the-visually-impaired-on-facebook/

Why do need Captioning?

- Image captioning closely related to **Visual Question Answering**
  -> If good descriptions of images can be generated. Then likely a variety of questions can be answered

- It seems there should be a strong relationship between **scene-graph generation** & **caption generation**.
  -> Then the benefits of SG generation we saw last time (image retrieval etc.) can be gained

- **Long term** goal perhaps **commentary** on a video frame-by-frame

"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

"boy is doing backflip on wakeboard."

"girl in pink dress is jumping in air."

"black and white dog jumps over bar."

"young girl in pink shirt is swinging on swing."

"man in blue wetsuit is surfing on wave."

Image credits: https://cs.stanford.edu/people/karpathy/sfmltalk.pdf

Datasets

COCO (C5)

1. A blue smart car parked in a parking lot.
2. Some vehicles on a very wet wide city street.
3. Several cars and a motorcycle are on a snow covered street.
4. Many vehicles drive down an icy street.
5. A small smart car driving in the city.

Flickr8K

1. A man is snowboarding over a structure on a snowy hill.
2. A snowboarder jumps through the air on a snowy hill.
3. a snowboarder wearing green pants doing a trick on a high bench
4. Someone in yellow pants is on a ramp over the snow.
5. The man is performing a trick on a snowboard high in the air.

Ref: http://cocodataset.org, https://illinois.edu/fb/sec/1713398

Key Goals in Image Captioning

- Automatic Caption generation
- Automatic caption evaluation

BLEU (2002)

Precision = 7/7

Modified precision= 2/7

Max . score is limited by freq. in reference.

Candidate	the	the	the	the	the	the	the
Reference 1	the	cat	is	on	the	mat	
Reference 2	there	is	a	cat	on	the	mat

Output is the geometric mean of n-gram score with a brevity penalty to discourage shorter translation.

Definition:

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-}gram' \in C'} Count(n\text{-}gram')}$$

Ref: Papineni, Kishore, et al. "BLEU: a method for automatic evaluation of machine translation." *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002.

Image credits: https://en.wikipedia.org/wiki/BLEU

ROUGE (2004)

Why bother about precision only?

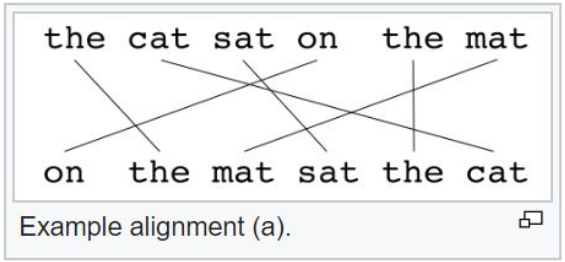
Originally developed as a package for evaluation of text summaries. Recall is used to encourage detailed description.

- ROUGE-N: N-gram recall between the candidate and the reference summaries.
- ROUGE-L: Longest Common Subsequence (LCS) based statistics.
- ROUGE-S: N-gram formation with skips.

Ref: Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." Text Summarization Branches Out (2004).

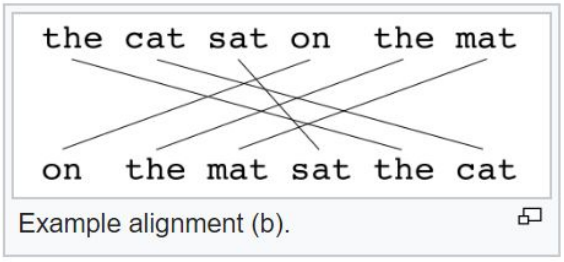
METEOR (2005)

It is based on an explicit word-to-word matching between the MT output being evaluated and one or more reference translations. It can also match synonyms.



Example alignment (a).

Calculate mapping between the candidate and reference caption. In conflict, mapping between least crosses is selected.



Example alignment (b).

$$F_{mean} = \frac{10PR}{R + 9P}$$

Extend it to longer n-grams with a pernaly for matching.

Ref: Banerjee, Satanjeev, and Alon Lavie. "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments." *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. 2005.
Image credits: https://en.wikipedia.org/wiki/METEOR

CIDEr (2014)

Use tf-idf metric to aggregate statistic for n-grams across the dataset.

Intuitively, words present across all caption is less informative, thus should be given less weight in evaluation of similarity.

tf ('this', D_1) = 1/7 tf ('this', D_1) = 2/7

Idf ('this',D) = log(2/2) = 0

tf-idf('this',D_1) = tf-idf('this',D_2) = 0

tf ('example', D_1) = 0 tf ('example', D_1) = 3/7

Idf ('example',D) = log(2/1) = 0.301

tf-idf('example',D_1) = 0 tf-idf('example',D_2) = 3/7 *0.301 = 0.13

Document 1

Term	Term Count
this	1
is	1
a	2
sample	1

Document 2

Term	Term Count
this	1
is	1
another	2
example	3

Ref: Vedantam, Ramakrishna, C. Lawrence Zitnick, and Devi Parikh. "Cider: Consensus-based image description evaluation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

Image credits: https://en.wikipedia.org/wiki/Tf–idf

Summary

1. Each of these metric is based on the n-gram matching.

2. N-gram overlap is neither necessary nor sufficient for the task for simulating human judgment in caption evaluation.

Exploring Nearest Neighbor Approaches for Image Captioning

Devlin, Jacob; Gupta, Saurabh; Girshick, Ross; Mitchell, Margaret; Zitnick, C Lawrence - 2015

Nearest-Neighbors Paper: Overview

- **Task:** how important is ***novel*** caption generation for automatic image captioning? (spoiler: quite important)
- **Benchmarked on:** MS COCO dataset
- **Compared to:** "From Captions to Visual Concepts and Back" [8]
- Prior papers had proposed idea: **find similar images => copy captions**
- **Motivation:**
  - Large datasets increase probability of finding appropriate caption
  - The work of Vinyals et al (2014) evidenced that *copying captions* may not be a terrible idea [35]
  - Find limitations of the very task of captioning
  - Explore properties of the largest caption dataset: MS COCO

Summary of the method

Given an uncaptioned query image **Q**, we would like to caption it:

(1) Find the **k** nearest neighbor images (NNs) in dataset

(2) Put the captions of all **k** images into a single set **C**

(3) Pick **c** in **C** with highest average lexical similarity over **C**

(4) **k** can be fairly large (50-200), so account for outliers during (3)

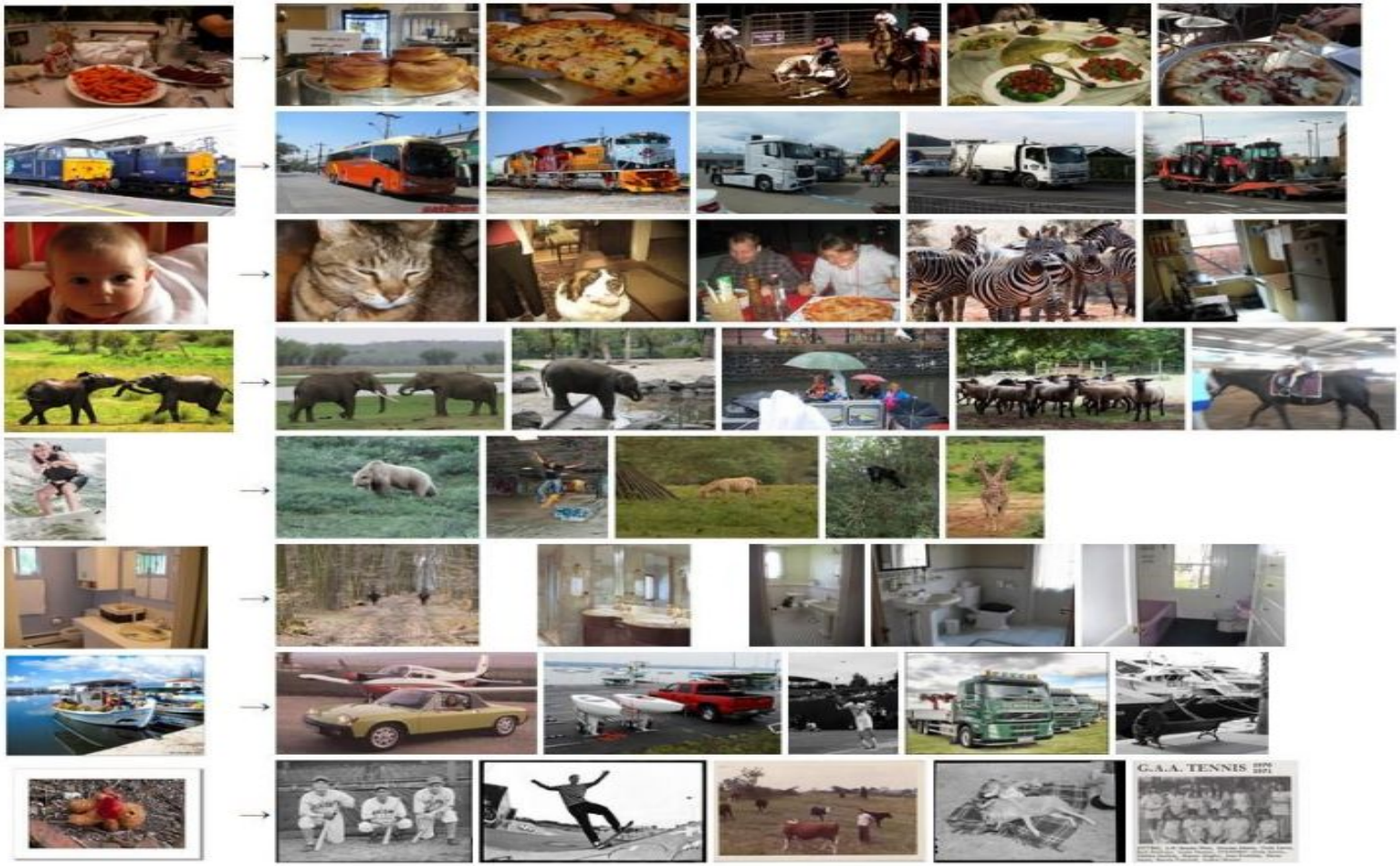
(5) Return **c** as the caption for **Q**

(1) Finding k Nearest-Neighbors

- Map **images** => **feature space**
- Three feature spaces experimented with:
  (i) GIST
  (ii) fc7
  (iii) fc7-fine
- Distance measure: **cosine similarity** of feature vectors

$$cosine(a, b) = \frac{\sum_{i=1}^{m} a_i b_i}{\sqrt{\sum_{i=1}^{m} a_i} \sqrt{\sum_{i=1}^{m} b_i}}$$

GIST



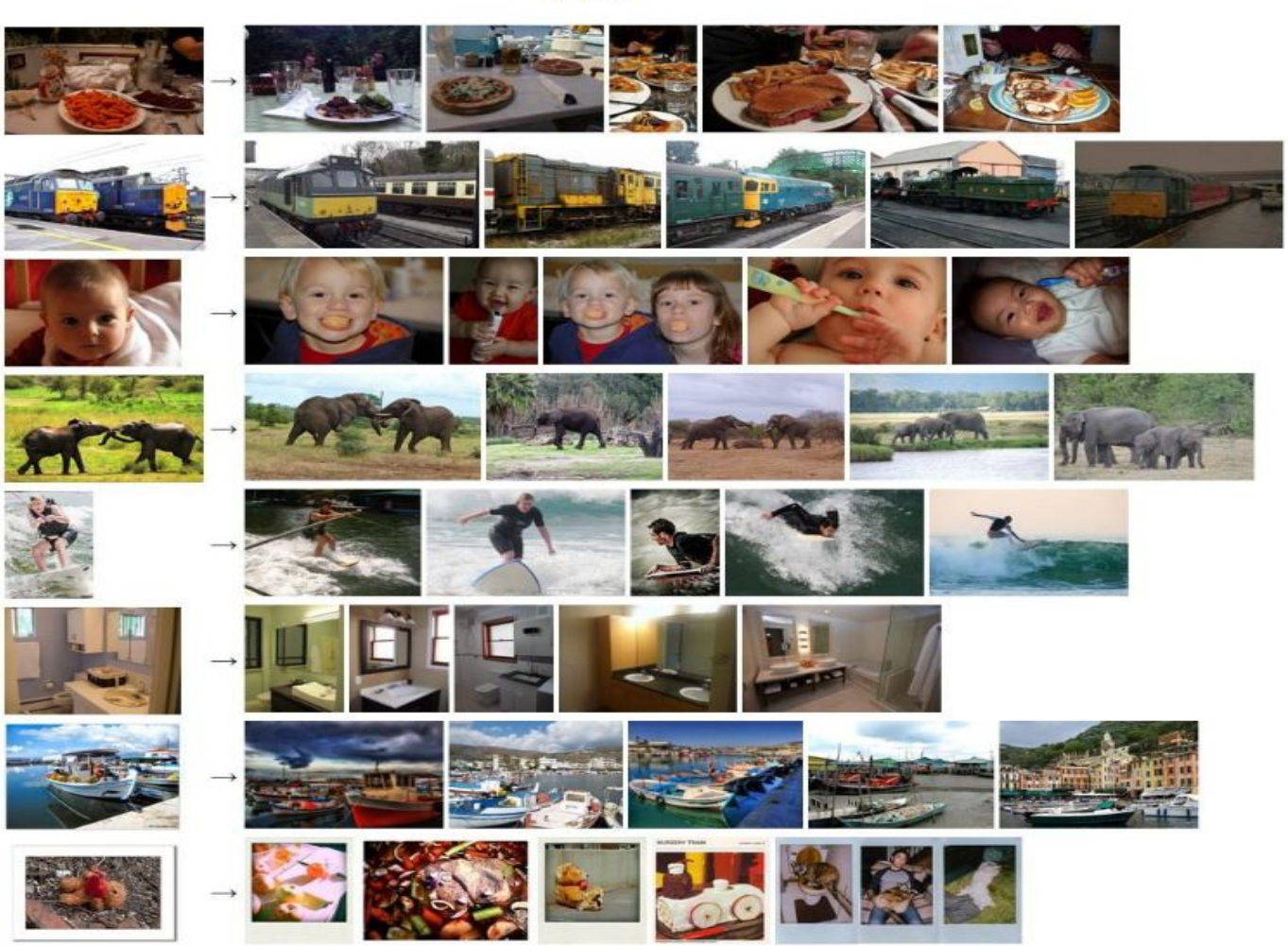
Reference:
Exploring Nearest Neighbor Approaches for Image Captioning - J Devlin et al (2015)

GIST

- A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. Progress in brain research, 2006)

- **Global** features based on **sum** of low level features (contours, textures..)
- Computed on images **resized** to 32 × 32 pixels (by the paper)

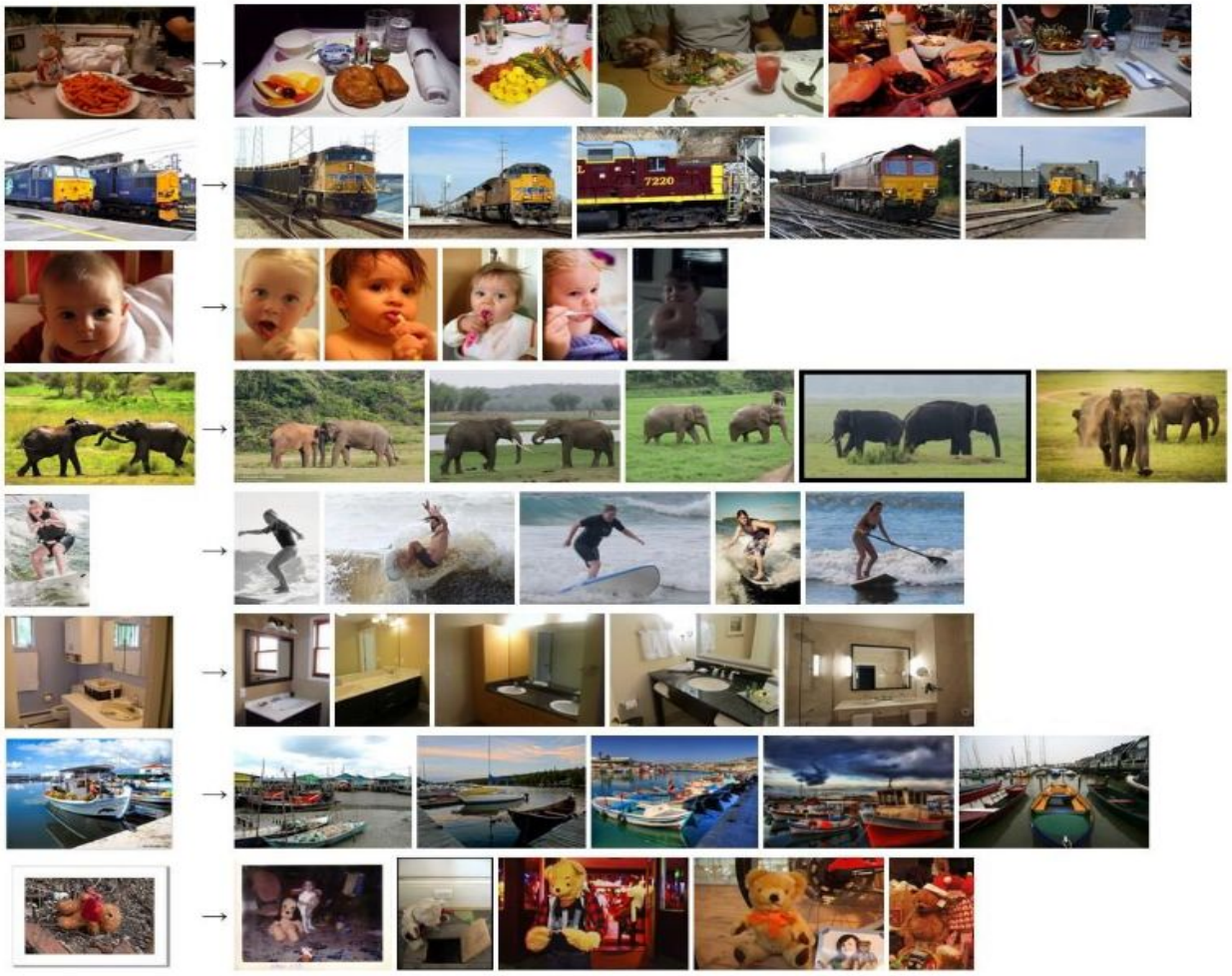
- GIST descriptor computation:
  1. Convolve image with **X** Gabor filters at **multiple** scales & orientations to produce 32 feature maps of input image size
  2. Divide each feature map into 16 regions (by a 4x4 grid), and then average the feature values within each region.
  3. Concatenate the 16 averaged values of all 32 feature maps, resulting in a 16x32=512 GIST descriptor.



- **deep CNN features** computed using **fc7** layer of the **VGG16 Net**
- Net trained on the 1000 ImageNet Classification Task

Reference:
Exploring Nearest Neighbor Approaches for Image Captioning - J Devlin et al (2015)

fc7-fine



- **fc7** layer again
- VGG weights initialized with ImageNet task
- **Fine-tuned** on image captioning task
- Classify the 1000 **most commonly** occuring words in image captions

Reference:
Exploring Nearest Neighbor Approaches for Image Captioning - J Devlin et al (2015)

(2),(3) Picking c* from caption set C

- **C** is union of all captions of the **k** Nearest Neighbors
- MS COCO has 5 captions/img so **|C| = 5k** (unless repeats)
- **Key idea**: pick the **consensus** caption **c*** in **C**

$$c^* = \underset{c \in C}{\operatorname{argmax}} \sum_{c' \in C} Sim(c, c')$$

- **Intuition:** a single caption in train data that describes **many** images **visually similar** to query image

Reference:
Exploring Nearest Neighbor Approaches for Image Captioning - J Devlin et al (2015)

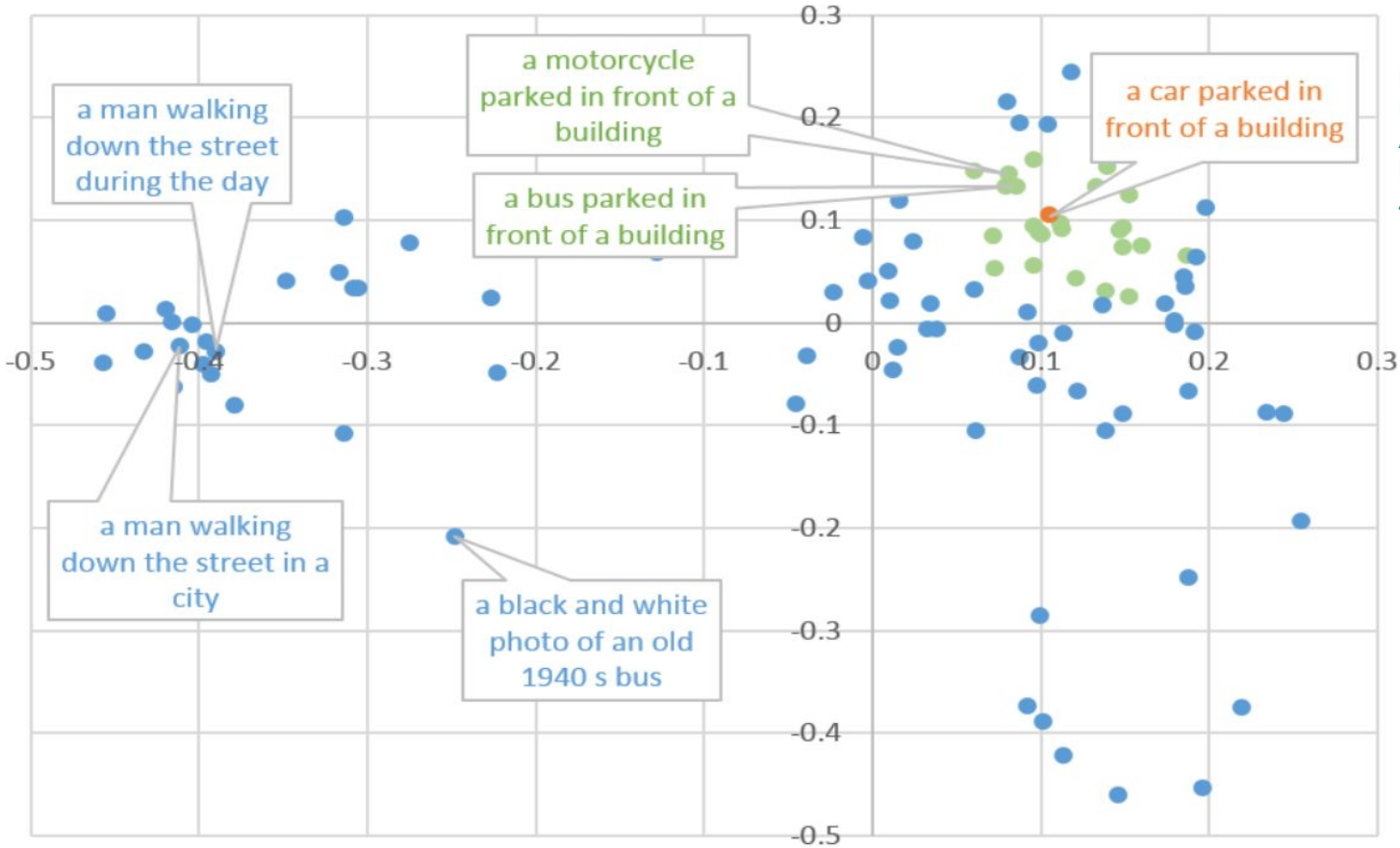
(4),(5) What of outliers in C?

- **k** is fairly large (50-200) so **|C|** is large (250-1000)
- For **robustness** to **outliers** & **noise**, use:

$$c^* = \underset{c \in C}{\operatorname{argmax}} \max_{M \subset C} \sum_{c' \in M} Sim(c, c')$$

- A second hyperparameter **m** introduced
- **Average similarity** over the **best** m-sized subset **M** of **C**
- **Intuition:** think of **c*** as the centroid of a large cluster of captions in **C** (want to find **best** such centroid)

Reference:
Exploring Nearest Neighbor Approaches for Image Captioning - J Devlin et al (2015)



Reference:
Exploring Nearest Neighbor Approaches for Image Captioning - J Devlin et al (2015)

Fig. 1: Example of the set of candidate captions for an image, the highest scoring m captions (green) and the consensus caption (orange). This is a real example visualized in two dimensions.

A point about the chosen c^*

J Devlin et al. make the following observation:

- If NN images are diverse, the chosen caption likely to be generic.
- If the NN images are quite similar, the chosen caption might end up being specific/descriptive.

Similarity (c,c') Measures

- Authors tried **BLEU** (1-to-4 gram overlap) & **CIDEr** (tf-idf weighted 1-to-4 gram overlap)
- Example captions from NNs approach on next slide
- **CIDEr** tends to favor more descriptive captions (likely due to preference for rarer **n-grams**)

Image	Selected Caption (BLEU)	Selected Caption (CIDEr)
	A bedroom with a bed and a couch.	A hotel room with two beds and a table.
	A train is stopped at a train station.	A red and white train parked in a train station.
	A group of people sitting around in a living room.	A group of people sitting on a couch in a living room.
	A group of people washing elephants in the water.	An elephant is swimming in the water near the rocks.
	Two zebras and a giraffe in a field.	Two zebras and a giraffe in a field.
	A car parked in front of a building.	A motorcycle parked in front of a brick building.
	A laptop computer sitting on top of a desk.	A laptop computer sitting on top of a desk.
	A clock sitting on top of a table.	A white airplane hanging from a ceiling in a museum.

Results & Evaluation

A point on evaluation

- **Train**, **validation** and **test** sets of **MS COCO**
- Split **validation** into 2: '**tuning**' & '**testval**'
- **testval** for experimenting with effects of:
  i) altering **k** and **m**
  ii) different feature spaces

- **test** set for reporting results on MS COCO

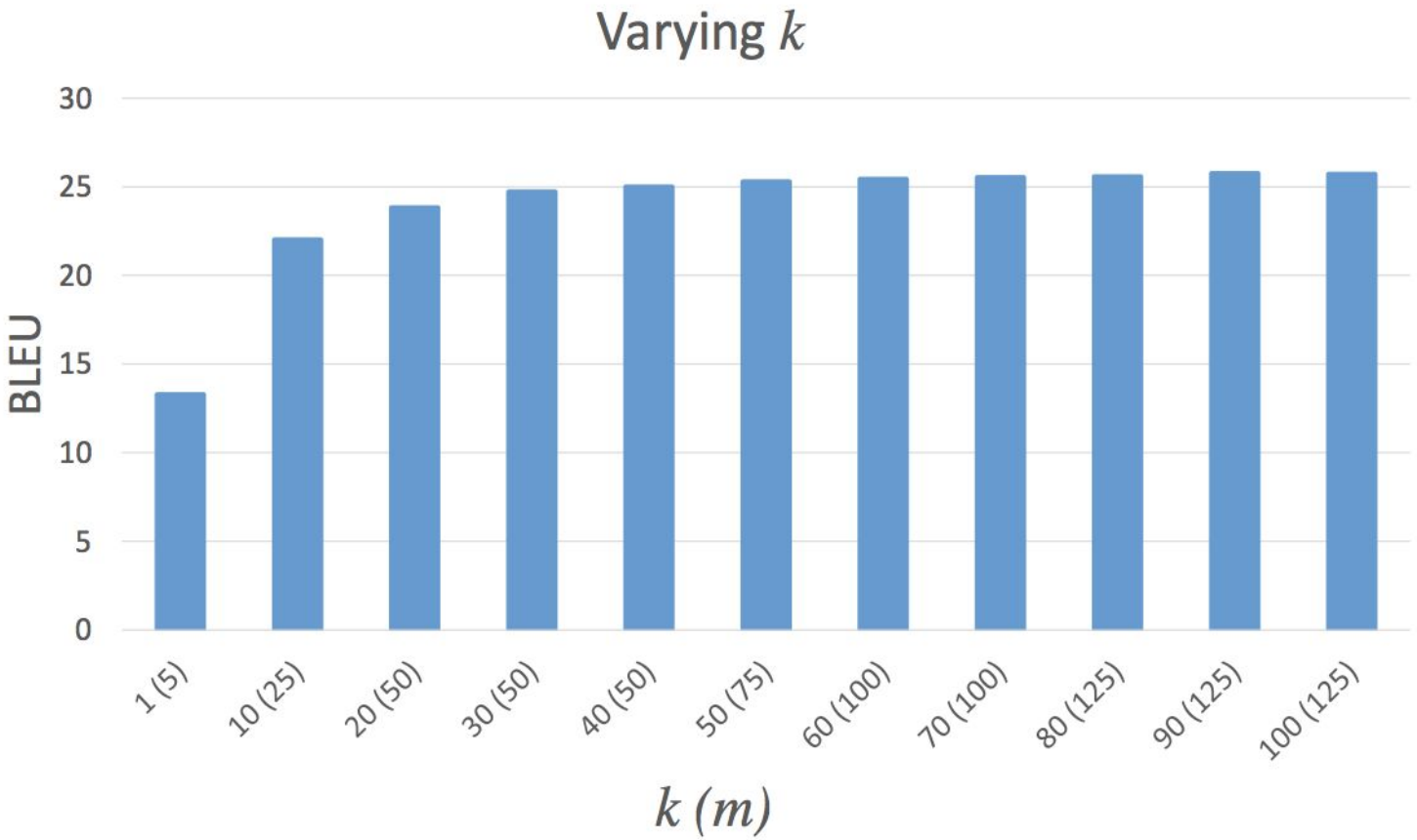


Fig. 3: Resulting BLEU scores when varying number of NN images, k. The optimal m for each k is shown in parentheses.

Reference:
Exploring Nearest Neighbor Approaches for Image Captioning - J Devlin et al (2015)

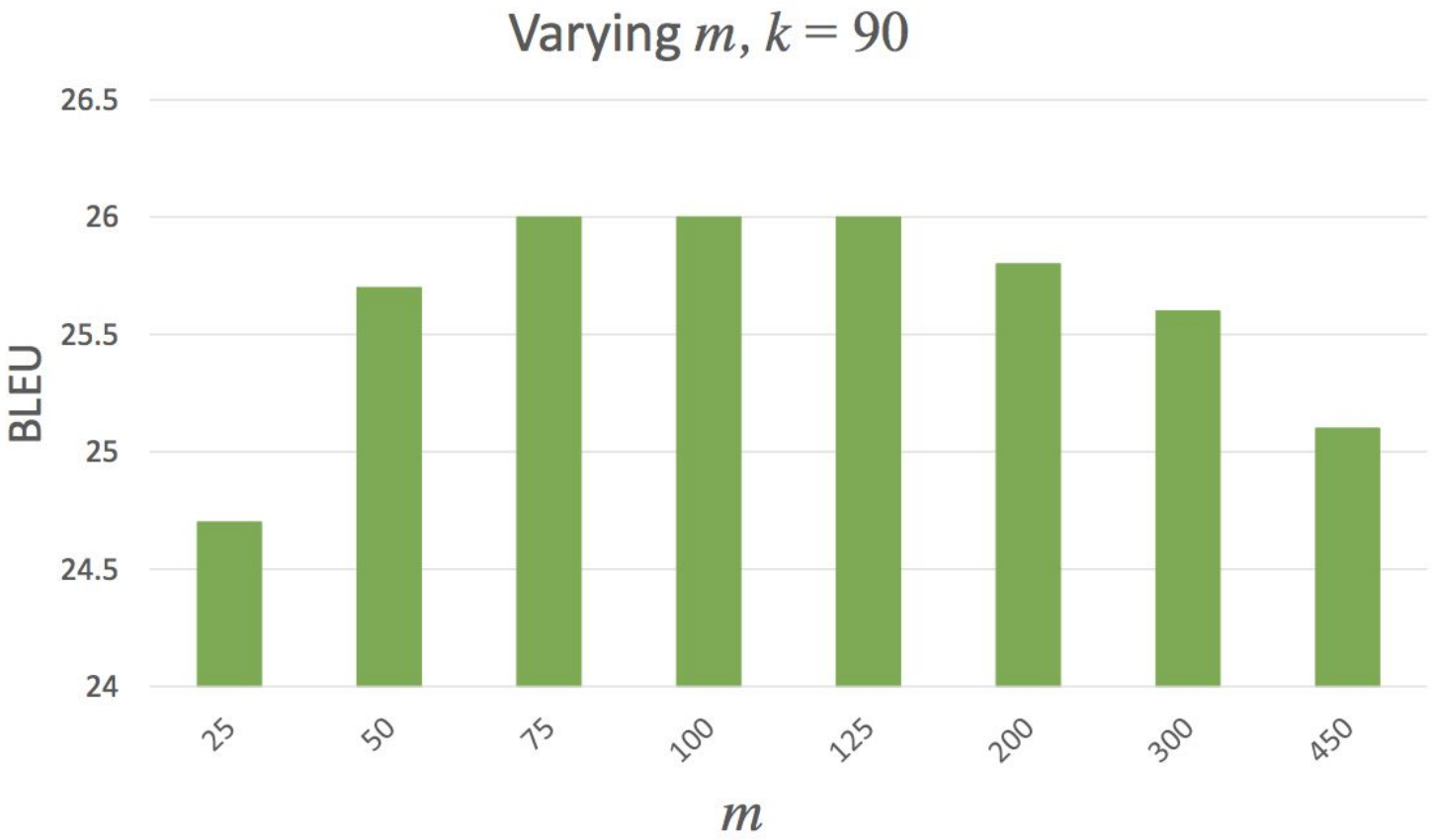


Fig. 4: Resulting BLEU scores when varying the number of captions m used to compute the consensus score. k is held constant at 90.

Reference:
Exploring Nearest Neighbor Approaches for Image Captioning - J Devlin et al (2015)

Comparing to [8]

Reference:
Exploring Nearest Neighbor Approaches for Image Captioning - J Devlin et al (2015)

Features	k	m	BLEU	CIDEr	METEOR
GIST	80	100	9.0	0.23	12.2
fc7	130	150	22.3	0.72	20.3
fc7-fine (BLEU)	90	125	26.0	0.85	22.5
fc7-fine (CIDEr)	80	200	25.1	0.90	22.8
ME + DMSM [8]			25.7	0.92	23.6

TABLE 2: BLEU [30], METEOR [4] and CIDEr [34] scores on testval for NN approaches using different feature spaces. See text for descriptions of the feature spaces.

On visual similarity with train data

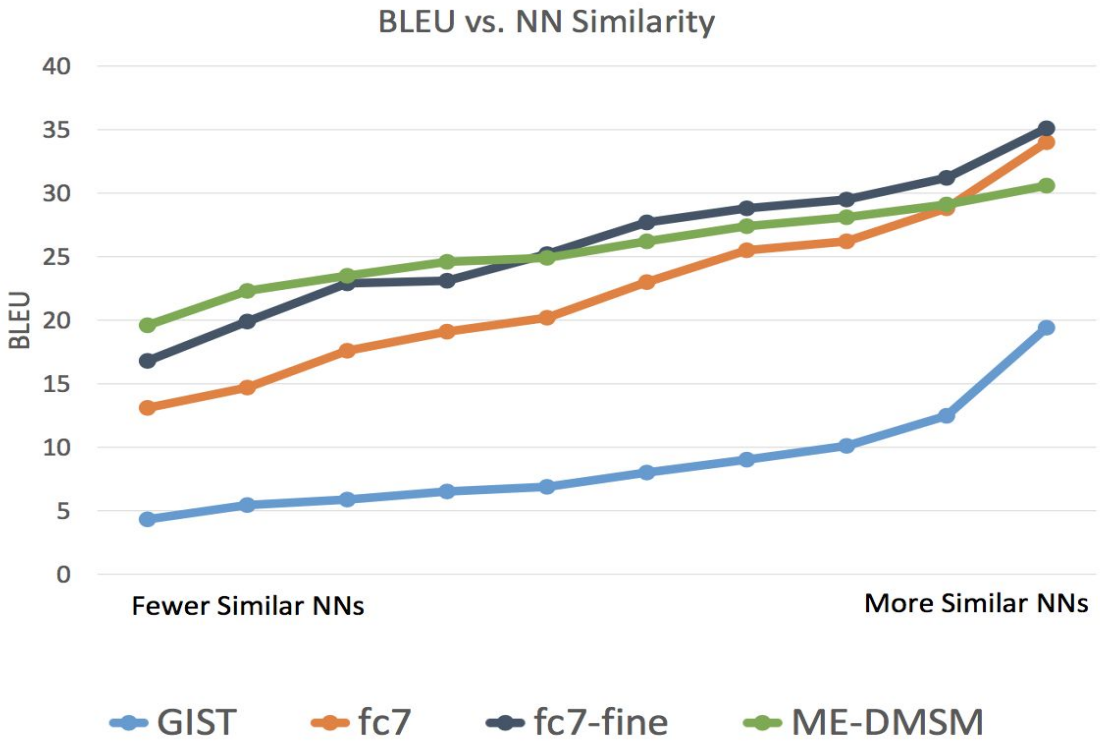


Fig. 5: BLEU scores for various approaches when the testval images are split into 10 equally sized bins based on visual similarity to the training data. The bins are arranged from those with fewer close NNs (left), to those with more NNs images (right).

Reference:
Exploring Nearest Neighbor Approaches for Image Captioning - J Devlin et al (2015)

Result on MS COCO test-set

Reference:
Exploring Nearest Neighbor Approaches for Image Captioning - J Devlin et al (2015)

Method	c5 BLEU 4	c5 CIDEr	c5 METEOR	c40 BLEU 4	c40 CIDEr	c40 METEOR
ME + DMSM [8]	**29.1**	**0.912**	24.7	**56.7**	**0.925**	33.1
LRCN [6]	27.7	0.869	24.2	53.4	0.891	32.2
Vinyals et al. [35]	27.2	0.834	23.6	53.8	0.842	32.7
Xu et al. [36]	26.8	0.850	24.3	52.3	0.878	32.3
m-RNN [25]	27.9	0.819	22.9	54.3	0.828	31.2
MLBL [18], [19]	26.0	0.740	21.9	51.7	0.752	29.4
NeuralTalk [16]	22.4	0.674	21.0	44.6	0.692	28.0
fc7-fine (CIDEr)	27.9 (2)	0.886 (2)	23.7 (3)	54.2 (2)	0.916 (2)	31.8 (5)
Human	21.7	0.854	**25.2**	47.1	0.910	**33.5**

Human evaluation

Reference:
Exploring Nearest Neighbor Approaches for Image Captioning - J Devlin et al (2015)

Approach	Human Judgements			BLEU
	Better	Equal	Better or Equal	
k-NN fc7-fine (BLEU)	5.5%	22.1%	27.6%	26.0
k-NN fc7-fine (CIDEr)	6.3%	20.2%	26.5%	25.1
ME + DMSM [8]	7.8%	26.2%	34.0%	25.7

TABLE 3: Results when comparing produced captions to those written by humans, as judged by humans. The percentage that are better than, equal to, and better than or equal to the captions written by humans are shown.

Thoughts on the paper

- Authors **don't** mention speed. NNs approaches are slow. Unsure how important that is?
- NNs does well implies caption datasets are too simplistic
- If captions **thoroughly** describe the image, NNs would likely be useless, and **novel** generation is necessitated
- Authors gives one way (Fig 5) to see how well approach generalizes to **semantically new** images
- Authors suggest research into **hybrid** (?) approaches might be useful

SPICE: Semantic Propositional Image Caption Evaluation

Peter Anderson, Basura Fernando, Mark Johnson, Stephen Gould - 2016

Slides adapted from: http://www.panderson.me/images/SPICE-slides.pdf

SPICE

- Motivation
  - Automatic caption evaluation
  - Drawbacks - existing metrics
- Key Ideas
  - Metric formulation
  - Comparison considering multiple captions
- Evaluation results

Captioning

The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.

Automate caption evaluation

The Evaluation Task: Given a candidate caption $\mathbf{c}_i$ and a set of **m** reference captions $R = \{r_1, \ldots, r_m\}$, compute a score ***S*** that represents similarity between c_i and ***R***.

c_1 Two women sitting at a white table next to a wall.

c_2 Photograph of small business store with white chairs and table.

Evaluation metric

"two women are sitting at a white table"

"two women sit at a table in a small store"

"two women sit across each other at a table smile for the photograph"

"two women sitting in a small store like business"

"two woman are sitting at a table"

R_i

Existing metrics

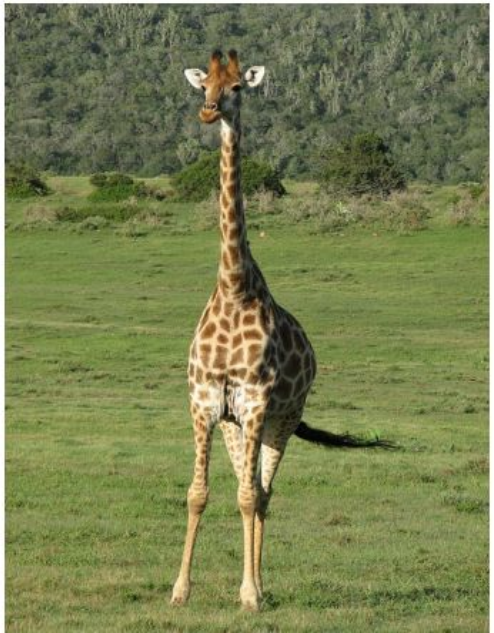
- BLEU
- METEOR
- ROGUE-L
- CIDEr

Based on n-gram similarity, which may not be sufficient to decide the best caption.

Why?

False Positives (High n-gram similarity)

A young girl standing on top of a tennis court.



A giraffe standing on top of a green field.

False Negatives (Low n-gram similarity)

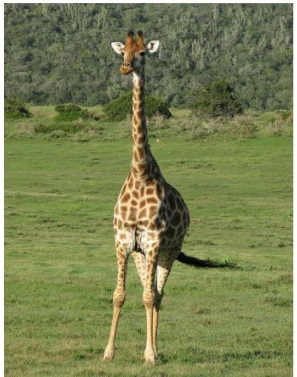
A shiny metal pot filled with some diced veggies.

The pan on the stove has chopped vegetables in it.

Note:: *SPICE isn't designed to avoid false negative.*

motivation

n-gram overlap is not necessary or sufficient for two sentences to mean the same.



State of the art

Is evaluation of image captions Solved?

	CIDEr-D	Meteor	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4
MSR Captivator[9]	0.937	0.339	0.68	0.907	0.819	0.71	0.601
Google[4]	0.946	0.346	0.682	0.895	0.802	0.694	0.587
m-RNN (Baidu/ UCLA)[16]	0.896	0.32	0.668	0.89	0.801	0.69	0.578
m-RNN[15]	0.935	0.325	0.666	0.89	0.798	0.687	0.575
MSR[8]	0.925	0.331	0.662	0.88	0.789	0.678	0.567
PicSOM[13]	0.856	0.318	0.654	0.875	0.775	0.663	0.554
Nearest Neighbor[11]	5	9	9	9	8	7	7
Berkeley LRCN[2]	0.891	0.322	0.656	0.871	0.772	0.653	0.534
Montreal/Toronto[10]	0.878	0.323	0.651	0.872	0.768	0.644	0.523
MLBL[7]	0.752	0.294	0.635	0.848	0.747	0.633	0.517
Tsinghua Bigeye[14]	0.682	0.273	0.616	0.866	0.756	0.628	0.493
ACVT[1]	0.716	0.288	0.617	0.831	0.713	0.589	0.478
Human[5]	6	3	11	6	12	12	13
NeuralTalk[12]	0.692	0.28	0.603	0.828	0.701	0.566	0.446
MIL[6]	0.69	0.284	0.596	0.827	0.707	0.564	0.432
Brno University[3]	0.536	0.252	0.509	0.716	0.541	0.392	0.278

source:: Lin Cui, Large-scale Scene UNderstanding Workshop, CVPR 2015

Related work

Semantic role labels:

Try to capture the the basic structure of the sentence -
'Who did what to whom, when, where and why'.

Sentence similarity is calculated by matching semantic frames across sentences by starting with the verbs at their head.

Verbs may not be meaningful/absent in a sentence.

A very tall building with a train sitting next to it.

Ref: Lo, C.k., Tumuluru, A.K., Wu, D.: Fully automatic semantic MT evaluation. In: ACL Seventh Workshop on Statistical Machine Translation. (2012)
Pradhan, S.S., Ward, W., Hacioglu, K., Martin, J.H., Jurafsky, D.: Shallow semantic parsing using support vector machines. In: HLT-NAACL. (2004) 233{240

SPICE

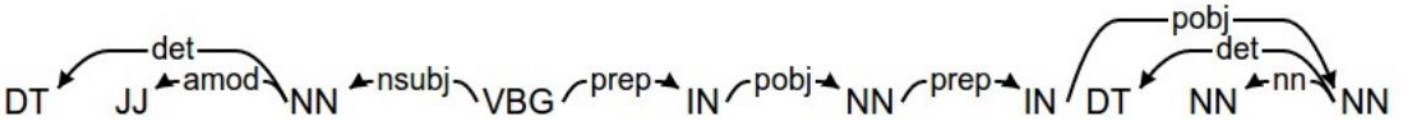
- Motivation
  - Automatic caption evaluation
  - Drawbacks - existing metrics

- Key Ideas
  - Metric formulation
  - Comparison considering multiple captions

- Experimental results

Scene Graphs

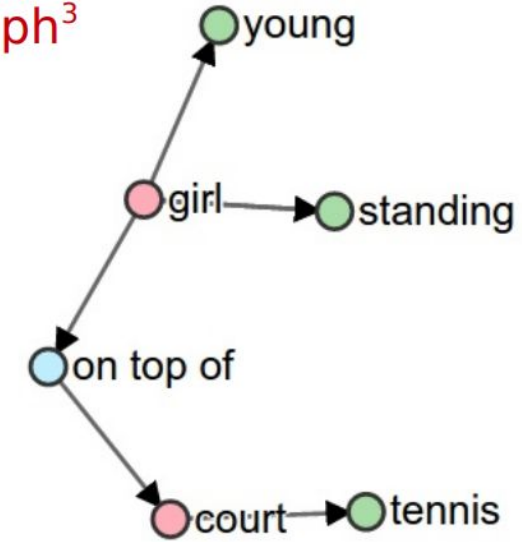
2. Parse[2]



1. Input

A young girl standing on top of a tennis court

3. Scene Graph[3]



4. Tuples

(girl)
(court)
(girl, young)
(girl, standing)
(court, tennis)
(girl, on-top-of, court)

Red : Object
Green: Attributes
Blue: Relations

2. Schuster, S., Krishna, R., Chang, A., Fei-Fei, L., Manning, C.D.: Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In: EMNLP 4th Workshop on Vision and Language. (2015)

3. Klein, D., Manning, C.D.: Accurate unlexicalized parsing. In: ACL. (2003)

"two women are sitting at a white table"

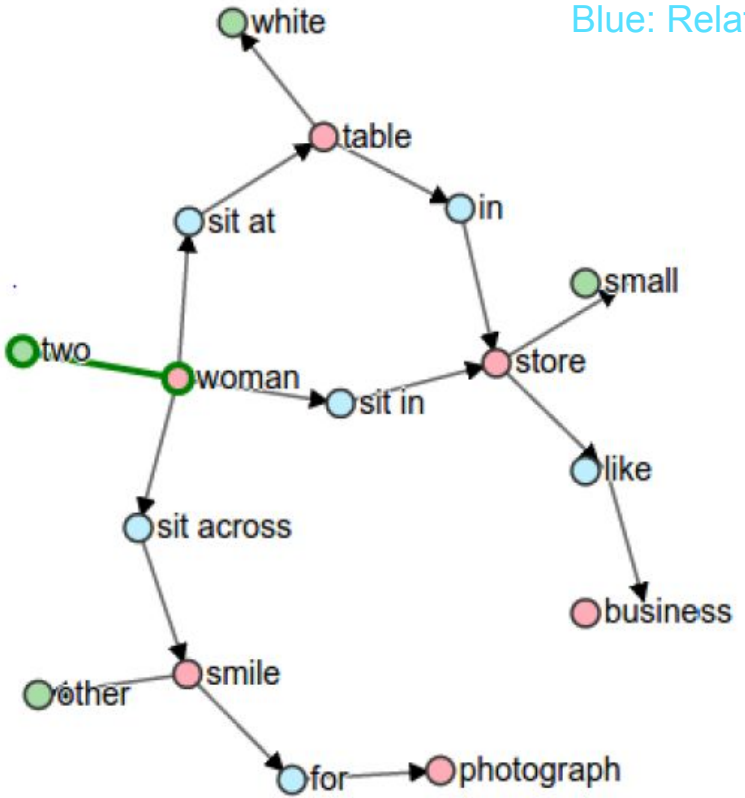
"two women sit at a table in a small store"

"two women sit across each other at a table smile for the photograph"

"two women sitting in a small store like business"

"two woman are sitting at a table"

Red : Object
Green: Attributes
Blue: Relations



SPICE calculation

Given candidate caption *c*, a set of reference captions *S*, and the mapping *T* from captions to tuples.

$$P(c, S) = \frac{T(c) \otimes T(S)}{T(c)}$$

$$R(c, S) = \frac{T(c) \otimes T(S)}{T(S)}$$

$$SPICE(c, S) = F_1(c, S) = \frac{2 * P(c, S) * R(c, S)}{P(c, S) + R(c, S)}$$

continued ...

SPICE calculated as an F-score over tuples. While matching tuples:

- Use wordnet synonym approach to consider tuples to be matched even if lemmatized forms are present.

- No partial credit even when only when element of the tuple is in correct.

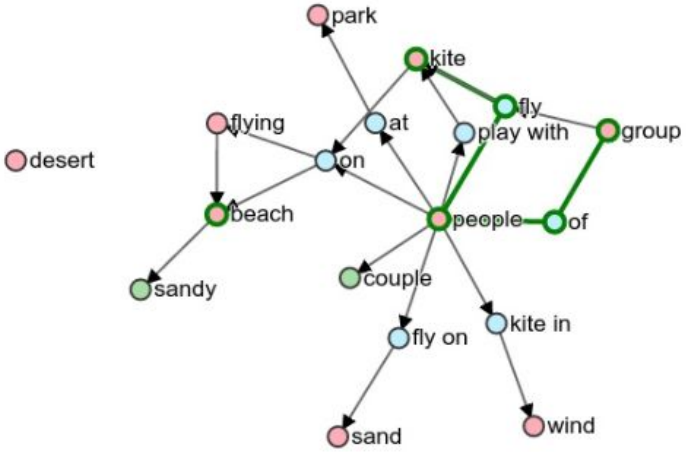
"*standing in park*" vs. "*sitting in park*" deserves no partial credit.

Good caption

Reference captions

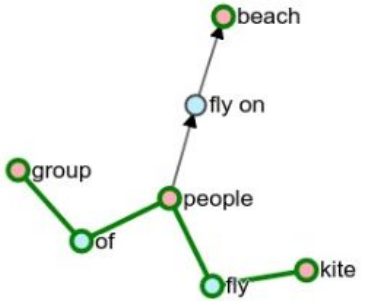
"People playing with kites outside in the desert."
"A group of people at a park flying a kite. "
"A group of people flying a kite on a sandy beach"
"People on the beach flying kites in the wind."
"A couple people out flying a kite on some sand."

Reference scene graph



Candidate caption & scene graph

"a group of people flying kites on a beach"



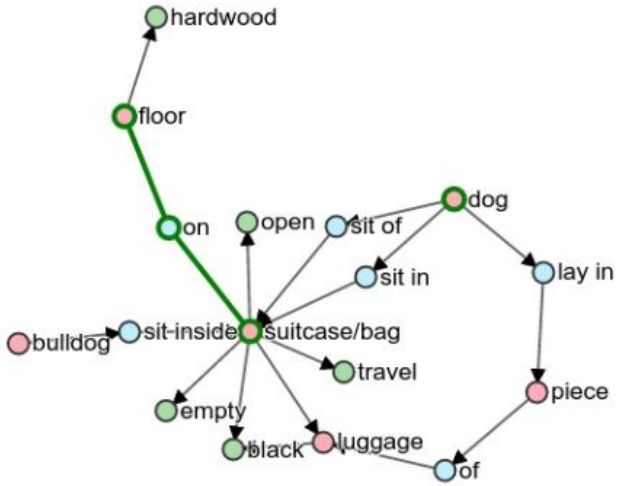
SPICE F-Score: 0.429, Pr: 0.857, Re: 0.286

Good caption

Reference captions

"a dog is sitting inside of a black suitcase"
"The bulldog is sitting inside the travel bag."
"A dog laying in a piece of black luggage."
"A dog sits in an open suitcase that is on a hardwood floor."
"A dog sitting inside an empty luggage bag on the floor"

Reference scene graph



Candidate caption & scene graph

"a dog sitting in a suitcase on the floor"

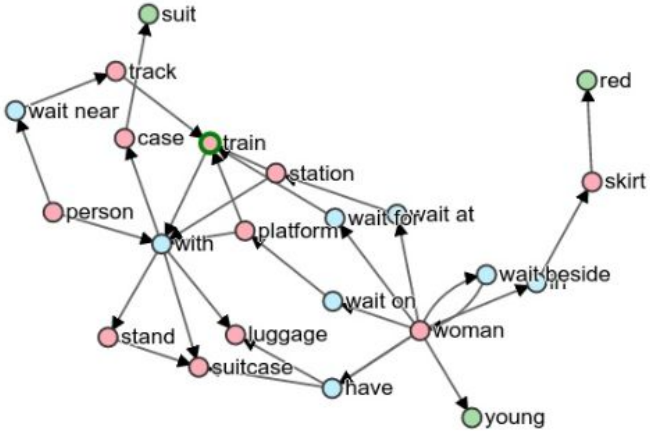
SPICE F-Score: 0.348, Pr: 1, Re: 0.211

Weak caption

Reference captions

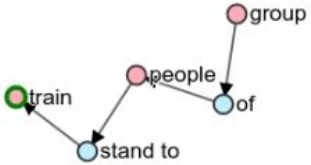
"A woman is waiting for a train. "
"A woman waiting at a train station with a suit case."
"A person with a suitcase stands waits near the train tracks. "
"A young woman in a red skirt is waiting on a train platform with her suitcase. "
"A woman waiting for a train with her luggage beside her."

Reference scene graph



Candidate caption & scene graph

"a group of people standing next to a train"



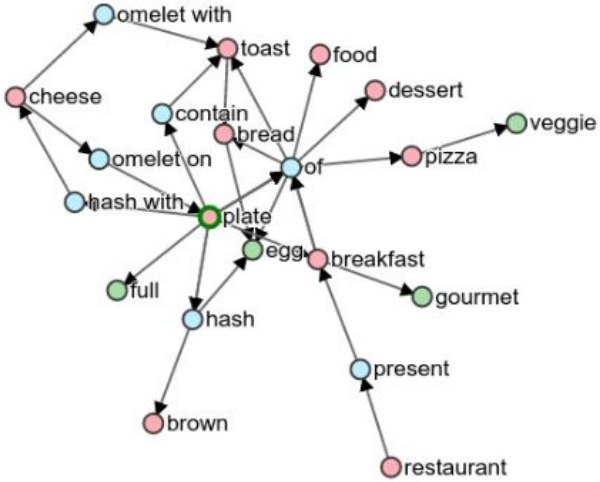
SPICE F-Score: 0.057, Pr: 0.2, Re: 0.033

Weak caption

Reference captions

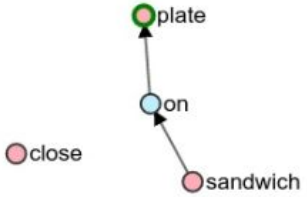
"The restaurant presents a gourmet breakfast of eggs and toast."
"A full plate of dessert, bread, and a veggie pizza. "
"A breakfast plate containing eggs, bread and french toast."
"A plate of food that includes toast, hash browns and eggs with cheese."
"A cheese omelet with toast on a plate."

Reference scene graph



Candidate caption & scene graph

"a close up of a sandwich on a plate"



SPICE F-Score: 0.059, Pr: 0.25, Re: 0.033

SPICE

- Motivation
  - Automatic caption evaluation
  - Drawbacks - existing metrics
- Key Ideas
  - Metric formulation
  - Comparison considering multiple captions
- Experimental results

	M1		M2		M3		M4		M5	
	ρ	p-value	ρ	p-value	ρ	p-value	ρ	p-value	ρ	p-value
Bleu-1	0.24	(0.369)	0.29	(0.271)	0.72	(0.002)	-0.54	(0.030)	0.44	(0.091)
Bleu-4	0.05	(0.862)	0.10	(0.703)	0.58	(0.018)	-0.63	(0.010)	0.30	(0.265)
ROUGE-L	0.15	(0.590)	0.20	(0.469)	0.65	(0.006)	-0.55	(0.030)	0.38	(0.142)
METEOR	0.53	(0.036)	0.57	(0.022)	0.86	(0.000)	-0.10	(0.710)	0.74	(0.001)
CIDEr	0.43	(0.097)	0.47	(0.070)	0.81	(0.000)	-0.21	(0.430)	0.65	(0.007)
SPICE-exact	0.84	(0.000)	0.86	(0.000)	0.90	(0.000)	0.39	(0.000)	0.95	(0.000)
SPICE	**0.88**	(0.000)	**0.89**	(0.000)	**0.89**	(0.000)	**0.46**	(0.070)	**0.97**	(0.000)

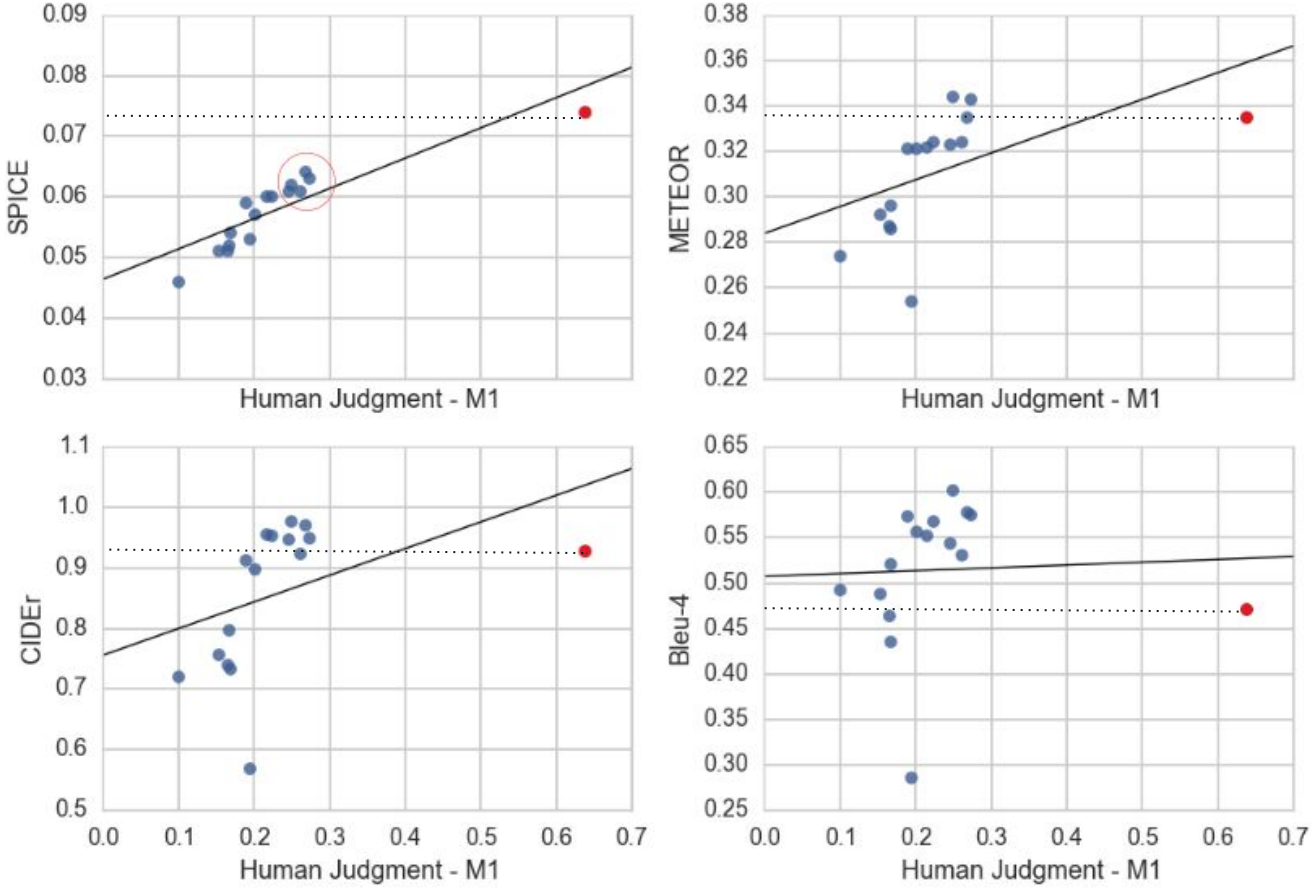
M1	Percentage of captions evaluated as better or equal to human caption.
M2	Percentage of captions that pass the Turing Test.
M3	Average correctness of the captions on a scale 1–5 (incorrect - correct).
M4	Average detail of the captions from 1–5 (lacking details - very detailed).
M5	Percentage of captions that are similar to human description.

System-level Pearson's correlation between evaluation metrics and human judgments for the 15 competition entries plus human captions in the 2015 COCO Captioning Challenge. SPICE more accurately reflects human judgment overall (M1-M2), and across each dimension of quality (M3-M5, representing correctness, detailedness and saliency)

Comparison

Absolute scores are lower with 40 reference captions (compared to 5 reference captions)

SPICE picks the same top-5 as human evaluators



	CIDEr-D	Meteor	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4
MSR Captivator[9]	0.937	0.339	0.68	0.907	0.819	0.71	0.601
Google[4]	0.946	0.346	0.682	0.895	0.802	0.694	0.587
m-RNN (Baidu/ UCLA)[16]	0.896	0.32	0.668	0.89	0.801	0.69	0.578
m-RNN[15]	0.935	0.325	0.666	0.89	0.798	0.687	0.575
MSR[8]	0.925	0.331	0.662	0.88	0.789	0.678	0.567
PicSOM[13]	0.856	0.318	0.654	0.875	0.775	0.663	0.554
Nearest Neighbor[11]	5	9	9	9	8	7	7
Berkeley LRCN[2]	0.891	0.322	0.656	0.871	0.772	0.653	0.534
Montreal/Toronto[10]	0.878	0.323	0.651	0.872	0.768	0.644	0.523
MLBL[7]	0.752	0.294	0.635	0.848	0.747	0.633	0.517
Tsinghua Bigeye[14]	0.682	0.273	0.616	0.866	0.756	0.628	0.493
ACVT[1]	0.716	0.288	0.617	0.831	0.713	0.589	0.478
Human[5]	6	3	11	6	12	12	13
NeuralTalk[12]	0.692	0.28	0.603	0.828	0.701	0.566	0.446
MIL[6]	0.69	0.284	0.596	0.827	0.707	0.564	0.432
Brno University[3]	0.536	0.252	0.509	0.716	0.541	0.392	0.278

source:: Lin Cui, Large-scale Scene UNderstanding Workshop, CVPR 2015

	SPICE	Object	Relation	Attribute	Color	Count	Size
Human [6]	**0.074**	**0.190**	**0.023**	**0.054**	0.055	**0.095**	**0.026**
MSR [38]	0.064	0.176	0.018	0.039	**0.063**	0.033	0.019
Google [39]	0.063	0.173	0.018	0.039	0.060	0.005	0.009
MSR Captivator [40]	0.062	0.174	0.019	0.032	0.054	0.008	0.009
Berkeley LRCN [1]	0.061	0.170	**0.023**	0.026	0.030	0.015	0.010
Montreal/Toronto [2]	0.061	0.171	**0.023**	0.026	0.023	0.002	0.010
m-RNN [41]	0.060	0.170	0.021	0.026	0.038	0.007	0.004
Nearest Neighbor [42]	0.060	0.168	0.022	0.026	0.027	0.014	0.013
m-RNN [43]	0.059	0.170	0.022	0.022	0.031	0.002	0.005
PicSOM	0.057	0.162	0.018	0.027	0.025	0.000	0.012
MIL	0.054	0.157	0.017	0.023	0.036	0.007	0.009
Brno University [44]	0.053	0.144	0.012	0.036	0.055	0.029	0.025
MLBL [45]	0.052	0.152	0.017	0.021	0.015	0.000	0.004
NeuralTalk [36]	0.051	0.153	0.018	0.016	0.013	0.000	0.007
ACVT	0.051	0.152	0.015	0.021	0.019	0.001	0.008
Tsinghua Bigeye	0.046	0.138	0.013	0.017	0.017	0.000	0.009
Random	0.008	0.029	0.000	0.000	0.000	0.004	0.000

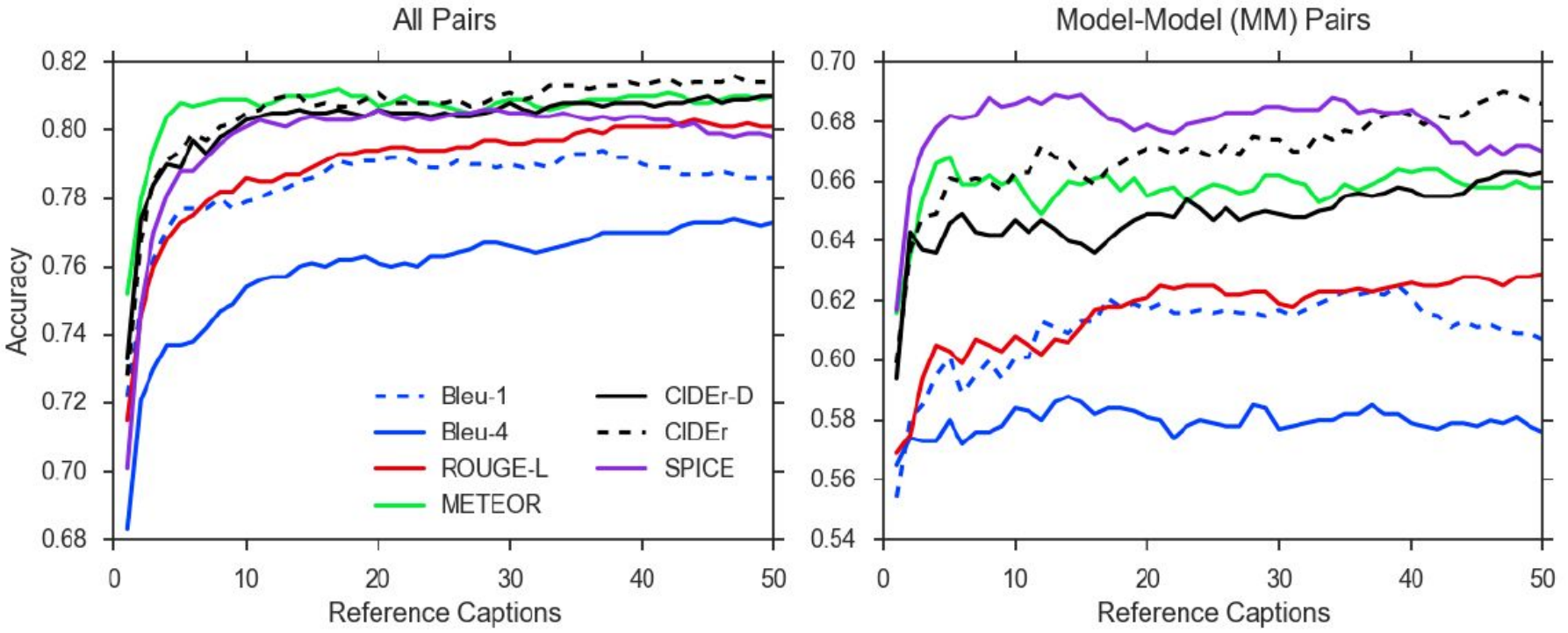
F-scores by semantic proposition subcategory. SPICE is comprised of object, relation and attribute tuples. Color, count and size are attribute subcategories. Although the best models outperform the human baseline in their use of object color attributes, none of the models exhibits a convincing ability to count.

	Flickr 8K [3]	Composite [35]
Bleu-1	0.32	0.26
Bleu-4	0.14	0.18
ROUGE-L	0.32	0.28
METEOR	0.42	0.35
CIDEr	0.44	0.36
SPICE	**0.45**	**0.39**
Inter-human	0.73	-

Caption-level Kendall's correlation between evaluation metrics and graded human quality scores. At the caption-level SPICE modestly outperforms existing metrics. All p-values (not shown) are less than 0.001

	HC	HI	HM	MM	All
Bleu-1	**64.9**	95.2	90.7	60.1	77.7
Bleu-2	56.6	93.0	87.2	58.0	73.7
ROUGE-L	61.7	95.3	91.7	60.3	77.3
METEOR	64.0	**98.1**	**94.2**	66.8	**80.8**
CIDEr	61.9	98.0	91.0	64.6	78.9
SPICE	63.3	96.3	87.5	**68.2**	78.8

Caption-level classification accuracy of evaluation metrics at matching human judgment on PASCAL-50S with 5 reference captions. SPICE is best at matching human judgments on pairs of model-generated captions (MM). METEOR is best at differentiating human and model captions (HM) and human captions where one is incorrect (HI). Bleu-1 performs best given two correct human captions (HC)



Pairwise classification accuracy of automated metrics at matching human judgment with 1-50 reference captions

Conclusion

- measures how well caption models recover objects, attributes and relations.

- Fluency is neglected. Include only objects, attributes and relations in the candidate caption for better score.

Room for improvement:

- Scene graph generation and evaluation
- Performance with large amount of reference captions
- Improved granularity.

;