Cross-validating Image Description Datasets and Evaluation Metrics

Josiah Wang & Robert Gaizauskas

Work presented as part of the chist-era D2K 2011 Visual Sense (ViSen) project

Generating Image Descriptions

A boy in black goggles leaping in the air by the beach.

Application: Information retrieval/indexing, for blind people to `see' image

Gold Standard Image Descriptions

A boy jumping in the air on the beach.

A boy with swimming trunks and goggles jumping on the sand by the beach.

A kid with sunglasses is jumping on the beach.

A young boy jumping in the air at the beach.

Boy in swim trunks jumping on beach.

Main Idea: Leave-one-out cross validation

Compare this description…

?

…against remaining gold reference descriptions for the same image.

A boy jumping in the air on the beach.
A boy with swimming trunks and goggles jumping on the sand by the beach.
A kid with sunglasses is jumping on the beach.
A young boy jumping in the air at the beach.
Boy in swim trunks jumping on beach.

Main Idea

Compare this description…

A boy jumping in the air on the beach.

…against remaining gold reference descriptions for the same image.

~~A boy jumping in the air on the beach.~~

A boy with swimming trunks and goggles jumping on the sand by the beach.

A kid with sunglasses is jumping on the beach.

A young boy jumping in the air at the beach.

Boy in swim trunks jumping on beach.

Main Idea

Repeat

Compare this description…

A boy with swimming trunks and goggles jumping on the sand by the beach.

…against remaining gold reference descriptions for the same image.

A boy jumping in the air on the beach.

~~A boy with swimming trunks and goggles jumping on the sand by the beach.~~

A kid with sunglasses in jumping on the beach.

A young boy jumping in the air at the beach.

Boy in swim trunks jumping on beach.

Main Idea

Repeat
for each

Compare this description…

A kid with sunglasses is jumping on the beach.

…against remaining gold reference descriptions for the same image.

A boy jumping in the air on the beach.

A boy with swimming trunks and goggles jumping on the sand by the beach.

~~A kid with sunglasses is jumping on the beach.~~

A young boy jumping in the air at the beach.

Boy in swim trunks jumping on beach.

Main Idea

Repeat for each reference

Compare this description…

A young boy jumping in the air at the beach.

…against remaining gold reference descriptions for the same image.

A boy jumping in the air on the beach.

A boy with swimming trunks and goggles jumping on the sand by the beach.

A kid with sunglasses in jumping on the beach.

~~A young boy jumping in the air at the beach.~~

Boy in swim trunks jumping on beach.

Main Idea

Repeat for each reference description

Compare this description…

Boy in swim trunks jumping on beach.

…against remaining gold reference descriptions for the same image.

A boy jumping in the air on the beach.

A boy with swimming trunks and goggles jumping on the sand by the beach.

A kid with sunglasses in jumping on the beach.

A young boy jumping in the air at the beach.

~~Boy in swim trunks jumping on beach.~~

Objectives

- Use leave-one-out cross validation to gain insights into:
  - Evaluation metrics
  - Image description datasets
- 'Bottom up' analysis
  - Human upper bound
  - Lower bound

Image Descriptions ≠ Image Captions

A man with red hair in a suit and a woman in a white dress and a crown on her head are cutting into a cake.

It's time for the cake cutting ceremony at my wedding! It was the most memorable day of my life!

Couples are increasingly spending more money for weddings, and this trend is predicted to continue for the next five years.

Image Description Datasets

- VLT2K
- UIUC PASCAL 1K
  - PASCAL50S
- Flickr30k
- Microsoft COCO
- ImageCLEF2015/2016
- Abstract Scenes
  - Abstract50S

UIUC PASCAL Sentences (PASCAL 1K)

A lone sheep walking through the woods.
A sheep in the morning mist with trees in the background.
A sheep standing on a hill at sunset.
a white sheep on the grass in front of trees
White sheep standing on grass in the morning.

Flickr 30K

A person hits a ball with a tennis racket.
A person swings at a tennis ball.
A tennis player wearing a green shirt about to hit a ball with his racquet.
A woman in a green shirt and blue hat is playing tennis.
Miami tennis player hits the ball with a forehand.

Microsoft COCO (MS COCO)

Soup salad and sandwich sitting on a plate.
A bowl of soup with a sandwich sits on a plate.
A meal of a salad soup and a sandwich.
A white plate topped with a bowl of soup next to a sandwich and salad.
A sandwich soup and salad all sit on a plate.

Visual & Linguistic Treebank (VLT2K)

A man is singing into a microphone. The rest of the band is also playing on stage.
A man is singing on stage with other men playing instruments. They are wearing t-shirts and it is dark in the background.
A young man belting out a song on the stage. A stage, bright lights and a microphone, with a group of students playing songs and singing as well.

Visual & Linguistic Treebank (VLT2K)

A man is singing into a microphone. ~~The rest of the band is also playing on stage.~~

A man is singing on stage with other men playing instruments. ~~They are wearing t-shirts and it is dark in the background.~~

A young man belting out a song on the stage. ~~A stage, bright lights and a microphone, with a group of students playing songs and singing as well.~~

ImageCLEF 2015/2016

A small eagle plushie against a white background.
A picture of a stuffed animal.
A stuffed bald eagle toy is sitting still.
A picture of a plush toy of a golden eagle, it has a white head and yellow beak, its body is entirely black and it's feet are also yellow.
Toy bird made of cloth with white head, black eyes, yellow beak and feet, and black body, wings and tail.

Abstract Scenes Dataset

Mike and Jenny are really cold.
The bear likes the fire.
The fire isn't keeping Mike and Jenny warm.

Set 1

The bear is looking at Mike and Jenny.
Mike and Jenny are afraid of the bear.
The bear is next to the campfire.

Set 2

Evaluation Metrics

- BLEU (from Machine Translation)
- ROUGE (from Summarization)
- Meteor (from Machine Translation)
- CIDEr (for Image Descriptions)

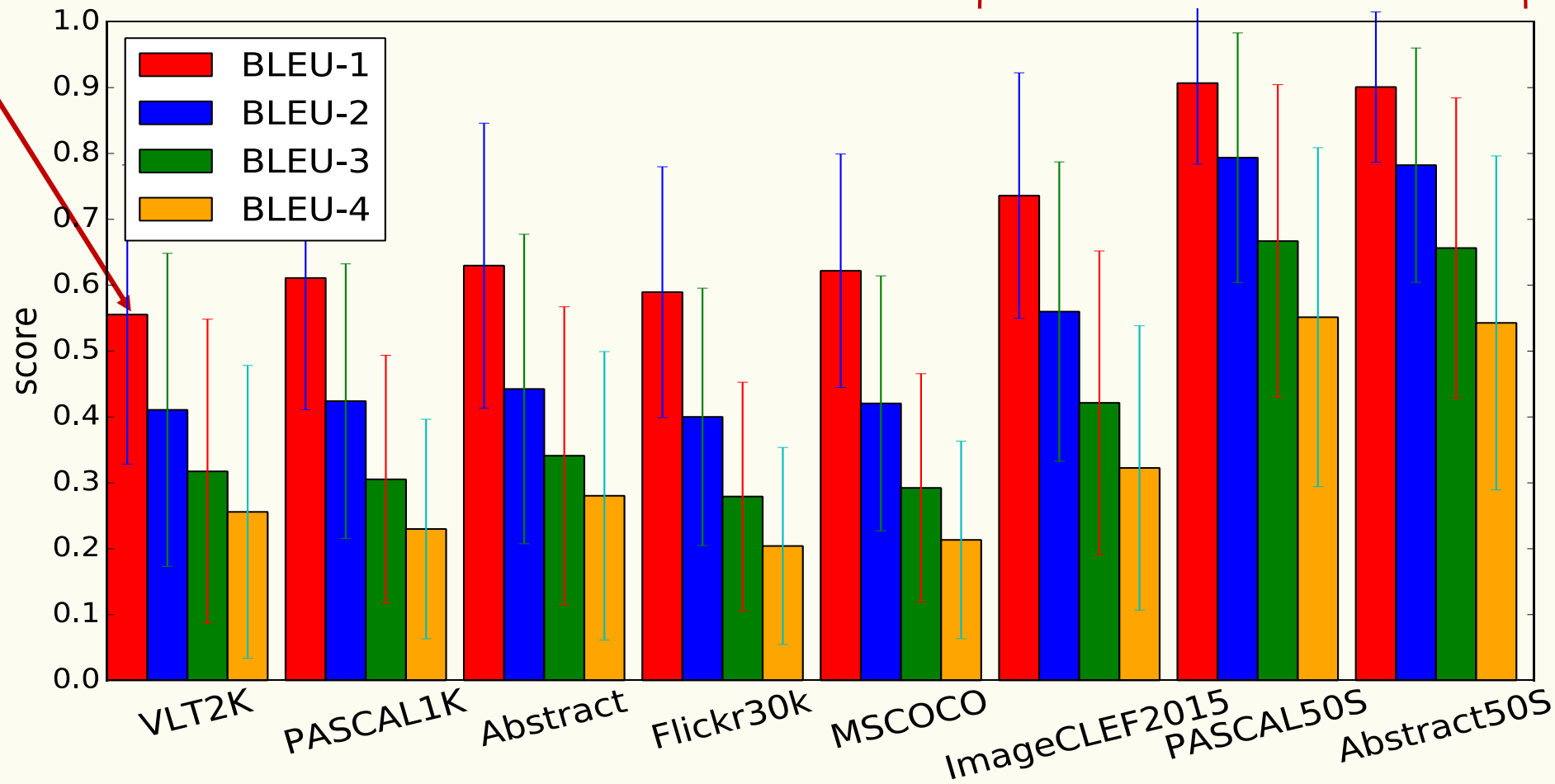
Human Upper-bound Evaluation

- Leave-one-out cross validation on reference descriptions of the same image
- Average scores per image, and then across whole dataset

BLEU

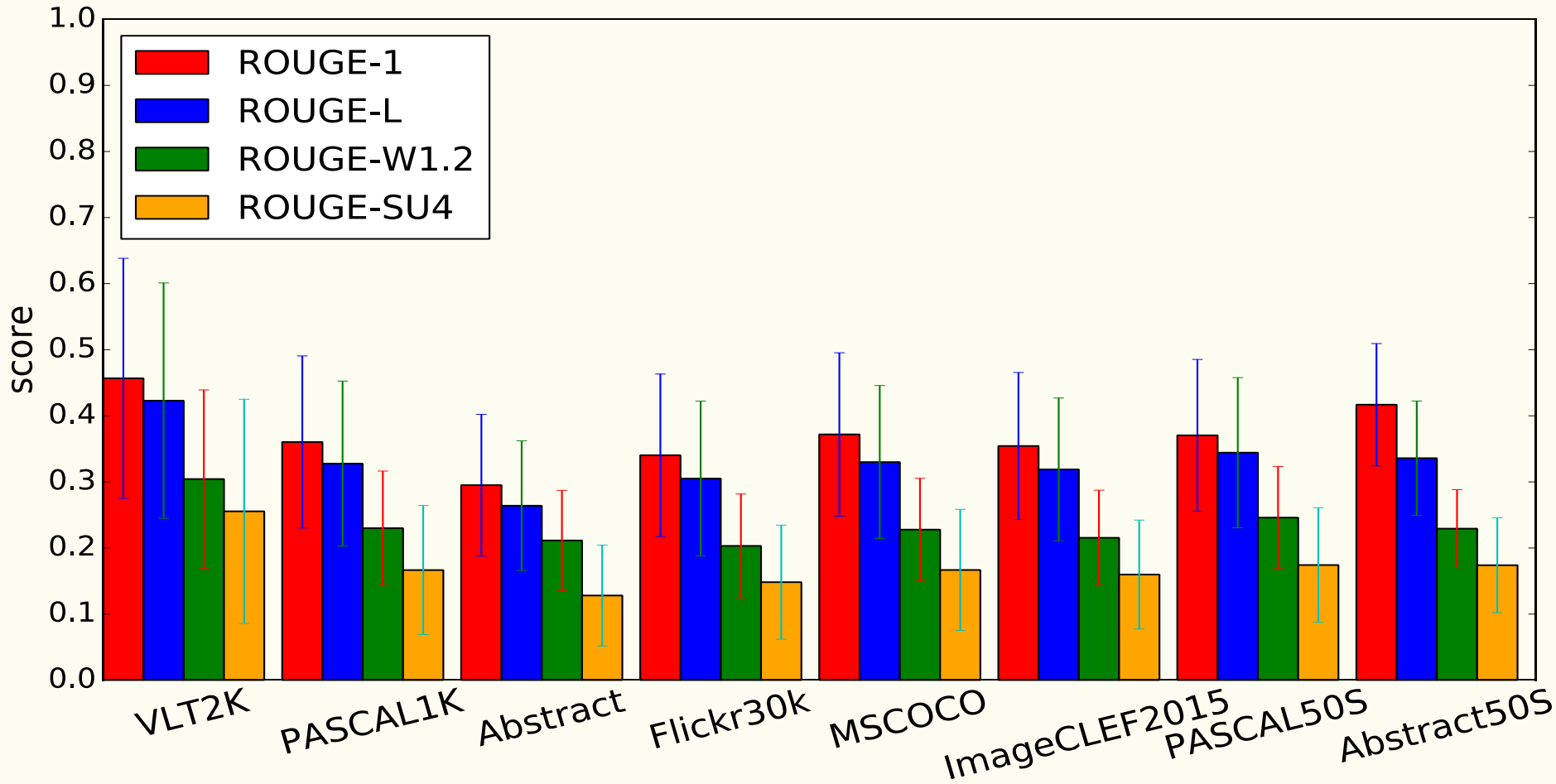
Low scores when few reference descriptions

High scores when many reference descriptions



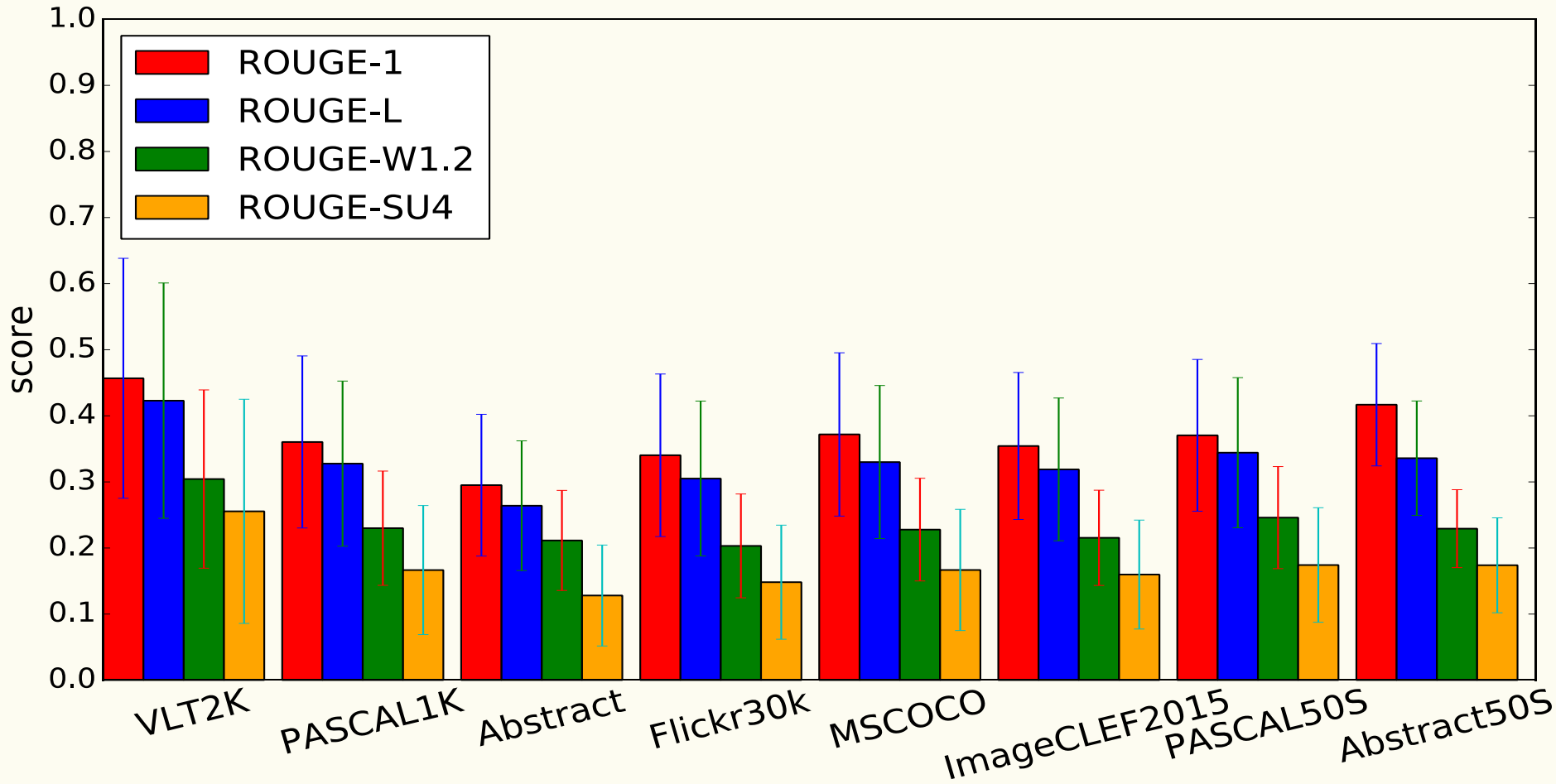
ROUGE

Absolute scores lower than BLEU …



ROUGE

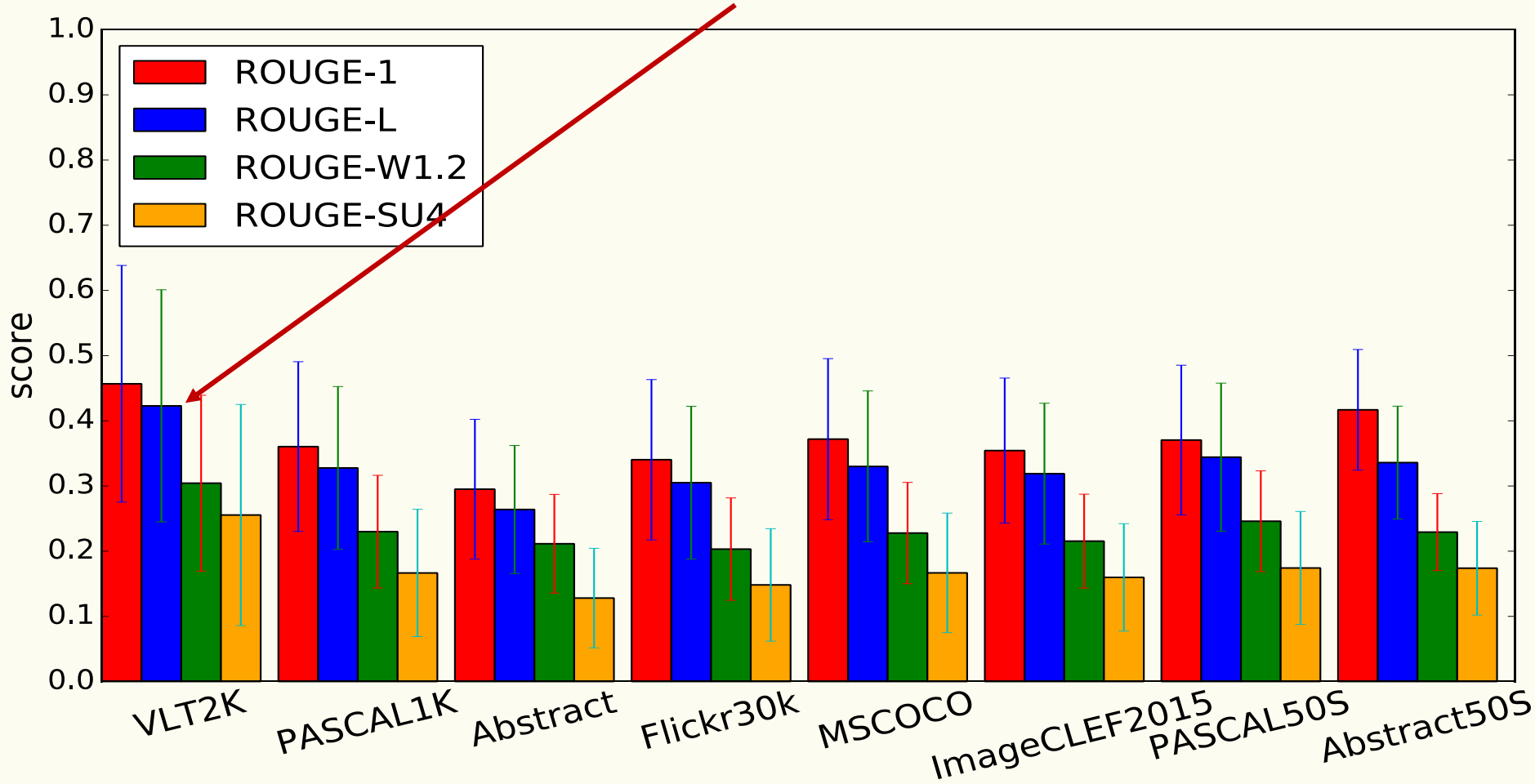
… but more uniform
(regardless of number of descriptions)



ROUGE

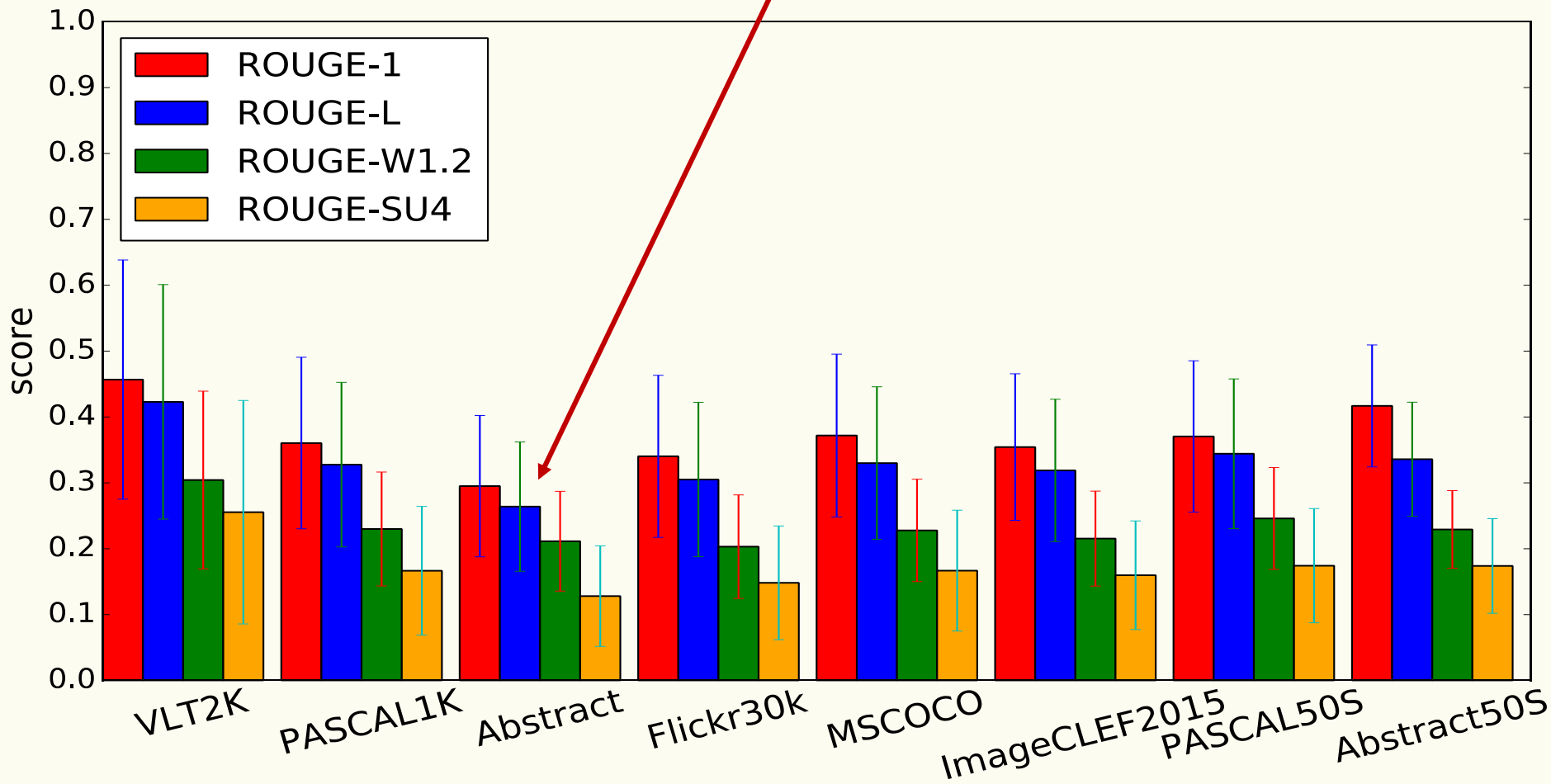
High mean score: constrained human actions

High standard deviation: could be described differently



ROUGE

Low scores – describe different aspects of scene



Abstract Scenes Dataset

Mike and Jenny are really cold.
The bear likes the fire.
The fire isn't keeping Mike and Jenny warm.

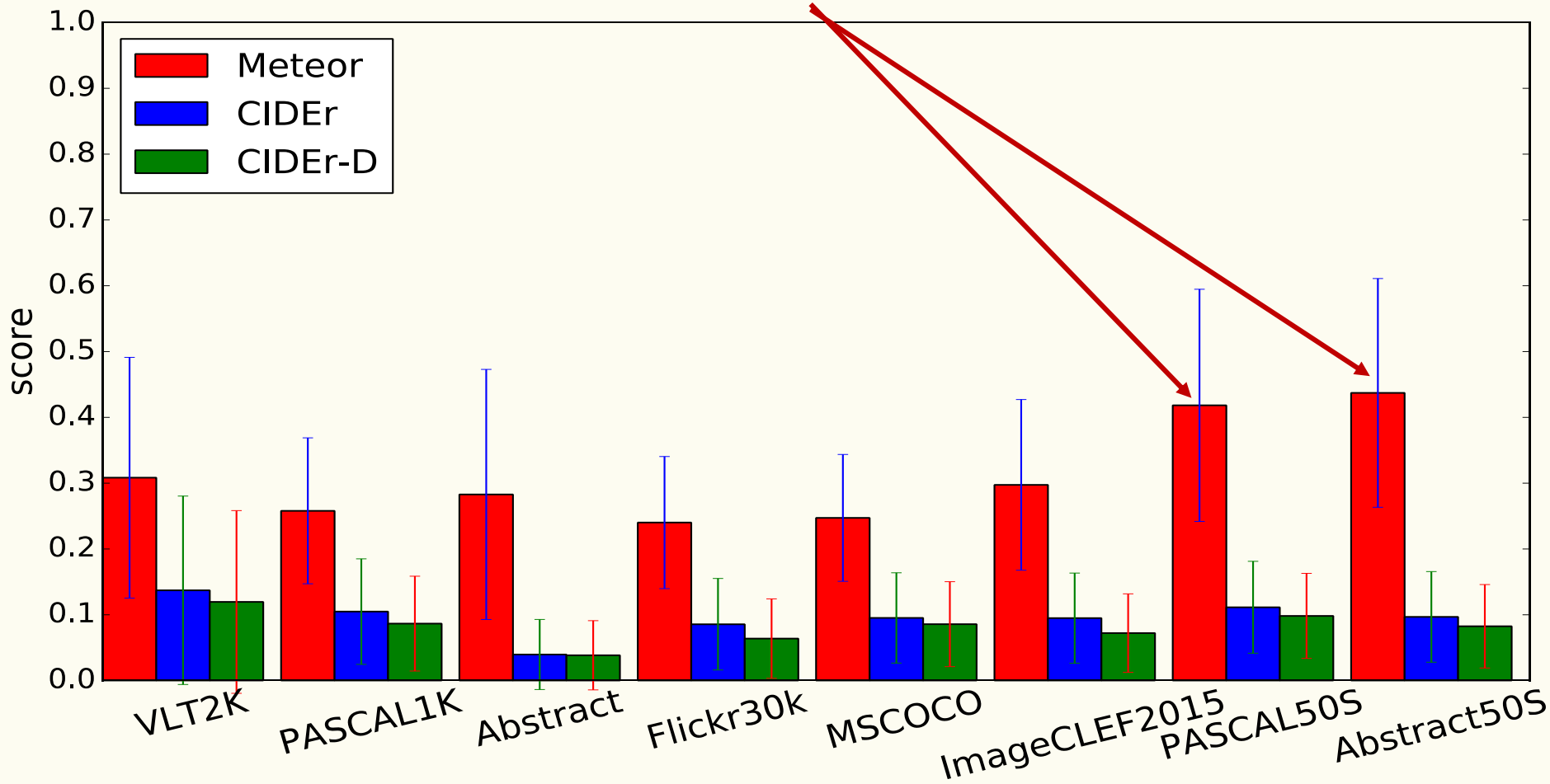
Set 1

The bear is looking at Mike and Jenny.
Mike and Jenny are afraid of the bear.
The bear is next to the campfire.

Set 2

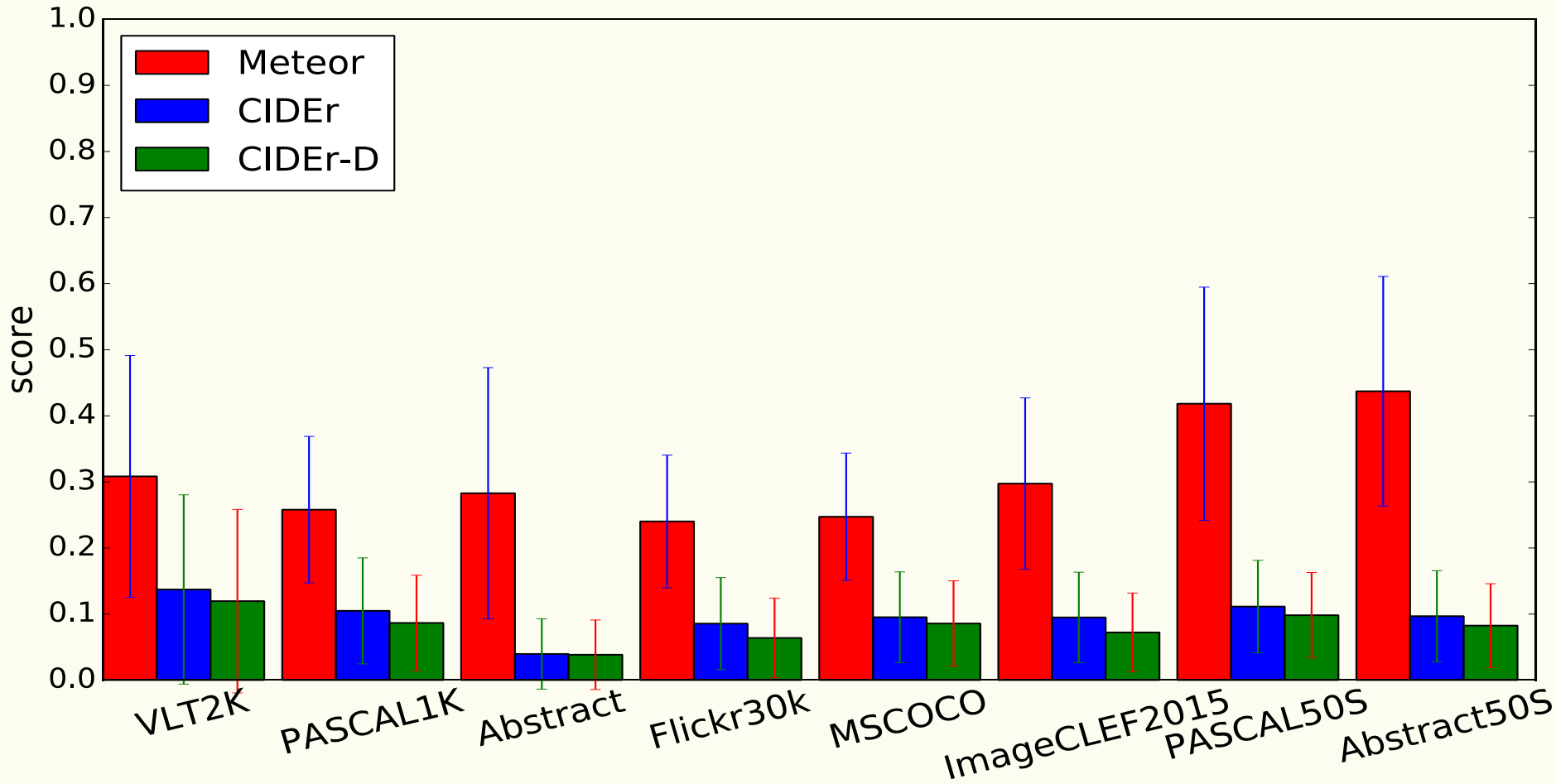
Meteor

Meteor: Quite dependent on number of reference descriptions



Meteor

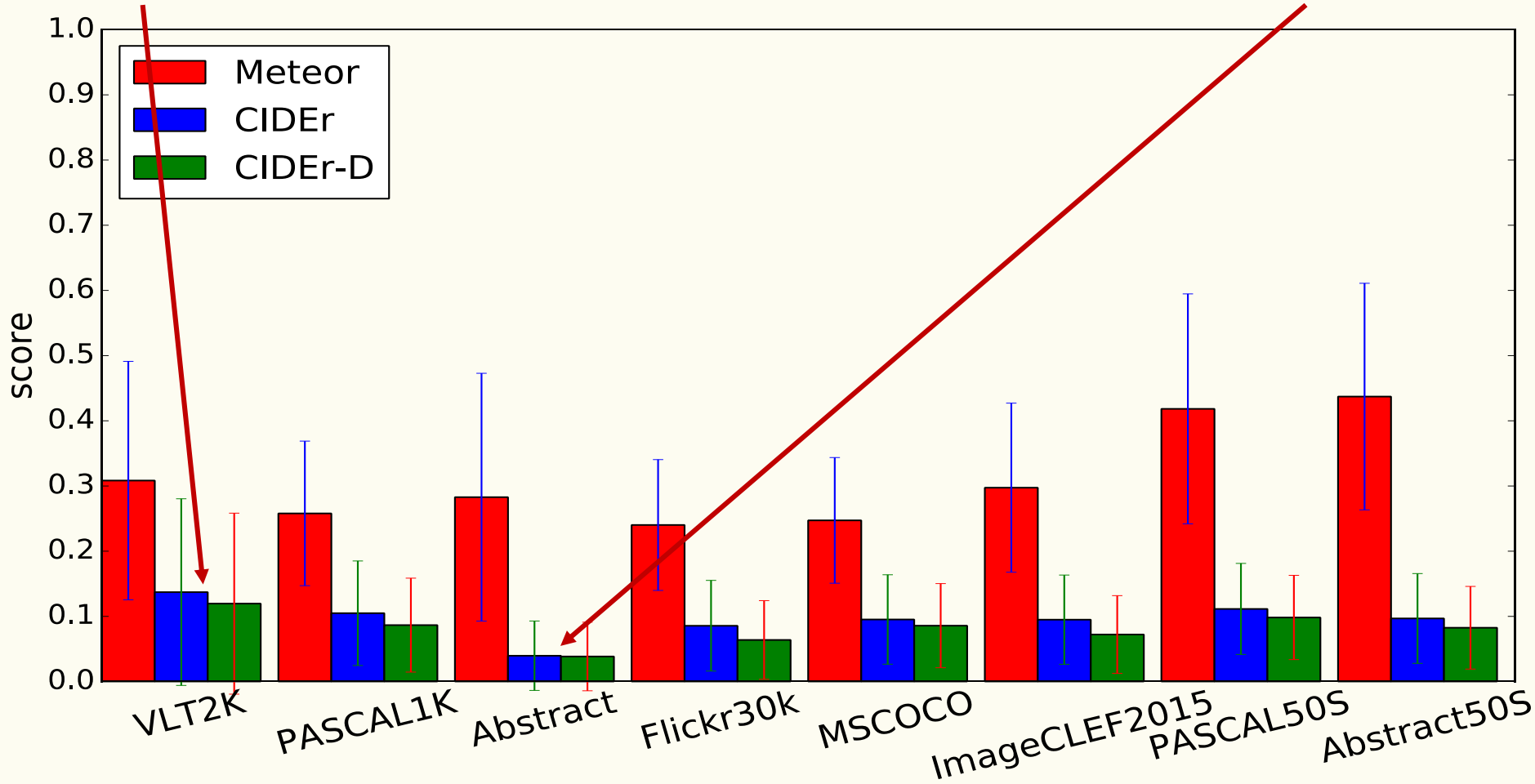
Meteor: Max score is 1.0 for all datasets
(at least one image has >= two identical descriptions)



CIDEr

Highest consensus (but also high variance)

Low consensus between descriptions



Lower bound Evaluation

- Upper bound to evaluate descriptions of same image
- Lower bound:
  - How much does descriptions vary within and across datasets?
  - How well does the metrics capture this?

Lower-bound Evaluation

Compare this description…

?

…against remaining gold reference descriptions for the same image.

A boy jumping in the air on the beach.
A boy with swimming trunks and goggles jumping on the sand by the beach.
A kid with sunglasses is jumping on the beach.
A young boy jumping in the air at the beach.
Boy in swim trunks jumping on beach.

Lower-bound Evaluation

Compare this description…

?

…against remaining gold reference descriptions for the same image.

~~A boy jumping in the air on the beach.~~

A boy with swimming trunks and goggles jumping on the sand by the beach.

A kid with sunglasses is jumping on the beach.

A young boy jumping in the air at the beach.

Boy in swim trunks jumping on beach.

Lower-bound Evaluation

From **another** random image from **same** dataset

Compare this description…

A man holds a chubby baby with pink cheeks and blue shirt.

…against remaining gold reference descriptions for the same image.

~~A boy jumping in the air on the beach.~~

A boy with swimming trunks and goggles jumping on the sand by the beach.

A kid with sunglasses is jumping on the beach.

A young boy jumping in the air at the beach.

Boy in swim trunks jumping on beach.

How similar are images in each dataset?

Lower-bound Evaluation

From **another** random image from **another** random dataset

Compare this description…

Mike and Jenny are upset about dropping the baseball.

…against remaining gold reference descriptions for the same image.

~~A boy jumping in the air on the beach.~~

A boy with swimming trunks and goggles jumping on the sand by the beach.

A kid with sunglasses is jumping on the beach.

A young boy jumping in the air at the beach.

Boy in swim trunks jumping on beach.

How domain-specific is the dataset?

Lower-bound Evaluation

From a random sentence from Brown corpus

Compare this description…

It is now a sweep of boulders and ledges with oak walnut and sumac creeping across the common and everywhere the ruins and the long long shadows.

…against remaining gold reference descriptions for the same image.

~~A boy jumping in the air on the beach.~~

A boy with swimming trunks and goggles jumping on the sand by the beach.

A kid with sunglasses is jumping on the beach.

A young boy jumping in the air at the beach.

Boy in swim trunks jumping on beach.

Make sure that metrics are measuring image descriptions!

Lower-bound Evaluation

From randomly generated `gibberish' from dataset vocab.

Compare this description…

and boat red while a station public down police coffee a biker

…against remaining gold reference descriptions for the same image.

~~A boy jumping in the air on the beach.~~

A boy with swimming trunks and goggles jumping on the sand by the beach.

A kid with sunglasses is jumping on the beach.

A young boy jumping in the air at the beach.

Boy in swim trunks jumping on beach.

How well does a metric evaluate structure?

Lower-bound Evaluation

From randomly generated `gibberish' from Brown corpus.

Compare this description…

he leadership such could the blow restaurant both hydrogen scattered the argue

…against remaining gold reference descriptions for the same image.

~~A boy jumping in the air on the beach.~~

A boy with swimming trunks and goggles jumping on the sand by the beach.

A kid with sunglasses is jumping on the beach.

A young boy jumping in the air at the beach.

Boy in swim trunks jumping on beach.

How well does a metric evaluate structure AND content?

Lower bound Evaluation: Summary

- BLEU
  - favours short sentences (precision)
  - doesn't capture structure well (even BLEU-4)
  - captures content fine

Lower bound Evaluation: Summary

- ROUGE
  - Dataset > Brown (domain specific)
  - Same Dataset > Different Dataset (dataset specific)
    - Especially Abstract and VLT2K
    - Different vocabularies, style etc.

Lower bound Evaluation: Summary

- Meteor
  - Captures dataset specificity even better than ROUGE
- CIDEr
  - Upperbound >> Random Intra dataset compared to other metrics

Discussion

- Proposed leave-one-out cross validation to gain insights into:
  - Image description datasets
  - Evaluation metrics
- Computed upper-bounds and lower-bounds
- Domain specific and dataset specific (esp. for Abstract Scenes)
- Future work:
  - "Top down" characterisation of datasets
  - Discovering which components are important for higher scores

Cross-validating Image Description Datasets and Evaluation Metrics

Josiah Wang & Robert Gaizauskas

Work presented as part of the chist-era D2K 2011 Visual Sense (ViSen) project