Overview of the ImageCLEF 2015 Scalable Image Annotation, Localization and Sentence Generation task

Andrew Gilbert, Luca Piras, Josiah Wang, Fei Yan, Emmanuel Dellandrea, Robert Gaizauskas, Mauricio Villegas and Krystian Mikolajczyk

Motivation & Aim

- Motivations:
  - Users struggle with the ever-increasing quantity of data available to them
  - Large number of images cheaply found and gathered from the Internet
  - More valuable is mixed modality data, web pages of both images and text

- Aim:
  - To develop techniques to allow computers to reliably describe images, localize the different concepts depicted in the images and generate a description of the scene, using noisy mixed modality data.

ImageCLEF 2015: Overview

- Single dataset of 500K webpages, images + text
- Subtask 1: Image localisation/detection
  - For each 500k image, annotate+localise with 251 concepts

- Subtask 2,3
  - Noisy track: Generate sentence for all 500k images
  - Clean track: Given bounding boxes (with concept labels) for 450 test images as input, generate sentences.

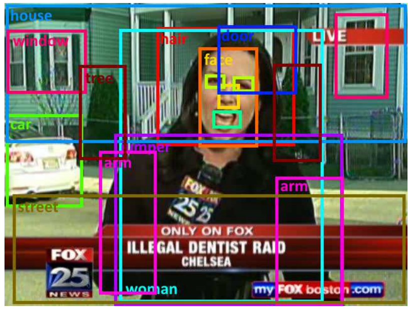
- Test dataset $\subset$ training dataset

(unknown to the user)

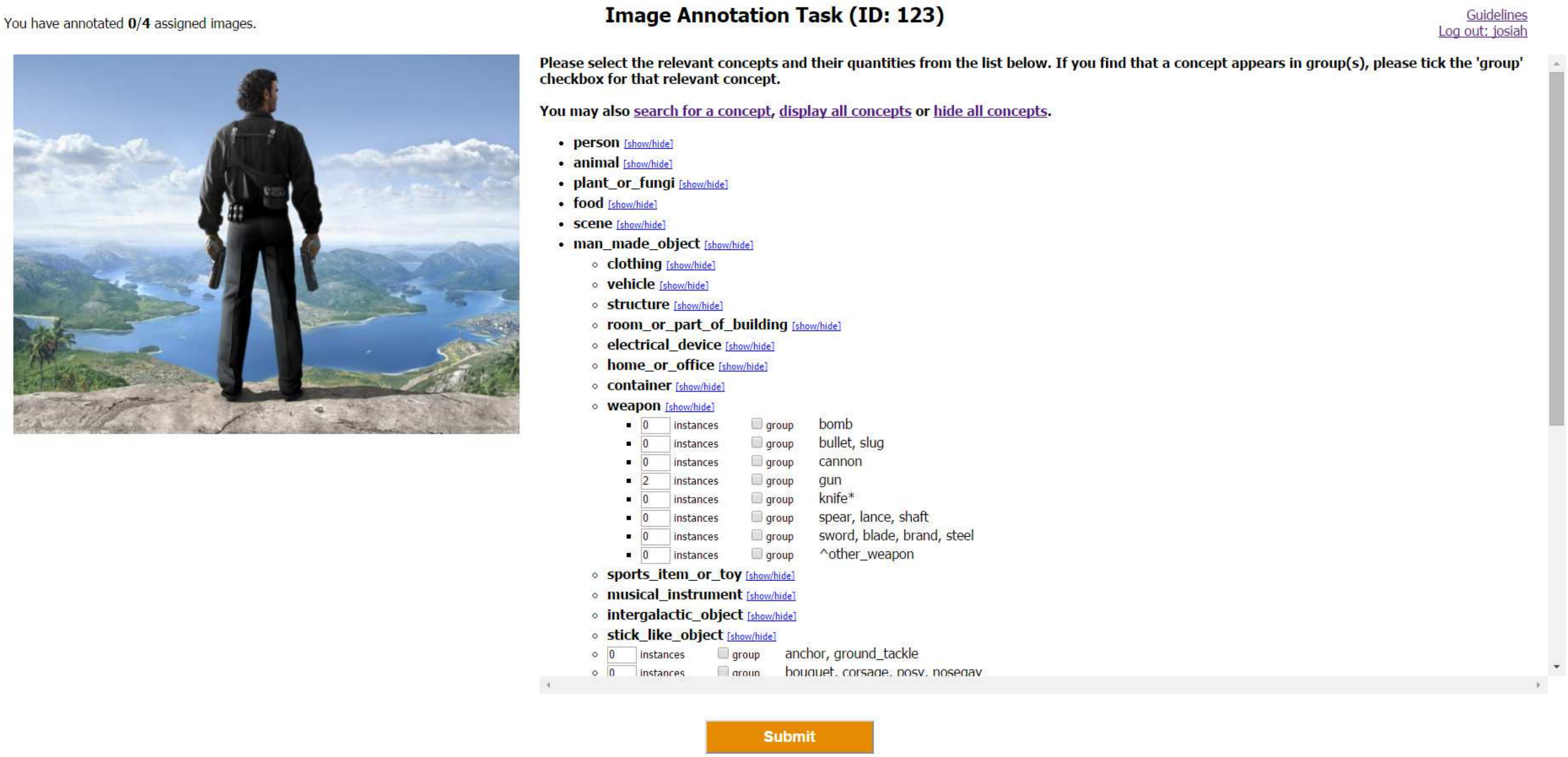
Training, development and test data

- 251 concepts from airplane to bottle to face & arm
  - Formed from looking at word co-occurence in 34M webpages of all English dictionary words
- Training/test Set of 500K images, >20 images per concept
  - CNN trained to identify “interesting images” for natural sentence generation

- The development set contained 1,979 images.
- Labelled Test of 3070 images (Sub task1, noisy2), 450 (clean2), both within the 500K
- Crowd sourced annotation of 5500 images
  - BBs of single instances or grouped instances
  - Annotations are not exhaustive



Crowd sourced Image level annotations



You have annotated **0/4** assigned images.

Image Annotation Task (ID: 123)

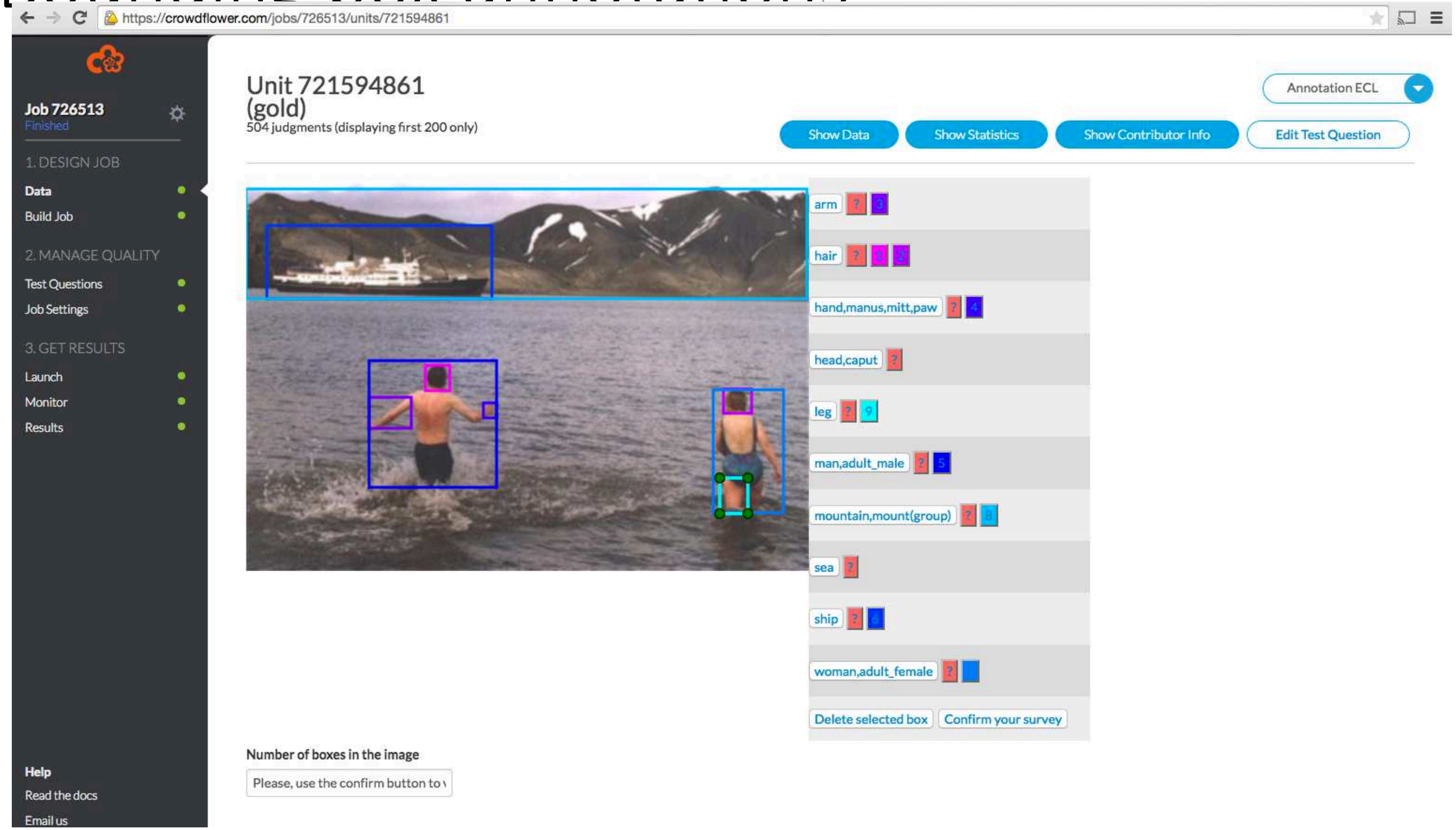
Guidelines
Log out: josiah

Please select the relevant concepts and their quantities from the list below. If you find that a concept appears in group(s), please tick the 'group' checkbox for that relevant concept.

You may also search for a concept, display all concepts or hide all concepts.

- **person** [show/hide]
- **animal** [show/hide]
- **plant_or_fungi** [show/hide]
- **food** [show/hide]
- **scene** [show/hide]
- **man_made_object** [show/hide]
  - **clothing** [show/hide]
  - **vehicle** [show/hide]
  - **structure** [show/hide]
  - **room_or_part_of_building** [show/hide]
  - **electrical_device** [show/hide]
  - **home_or_office** [show/hide]
  - **container** [show/hide]
  - **weapon** [show/hide]
    - 0 instances □ group bomb
    - 0 instances □ group bullet, slug
    - 0 instances □ group cannon
    - 2 instances □ group gun
    - 0 instances □ group knife*
    - 0 instances □ group spear, lance, shaft
    - 0 instances □ group sword, blade, brand, steel
    - 0 instances □ group ^other_weapon
  - **sports_item_or_toy** [show/hide]
  - **musical_instrument** [show/hide]
  - **intergalactic_object** [show/hide]
  - **stick_like_object** [show/hide]
  - 0 instances □ group anchor, ground_tackle
  - 0 instances □ group bouquet, corsage, posy, nosegay

Bounding box annotations



https://crowdflower.com/jobs/726513/units/721594861

Unit 721594861 (gold)

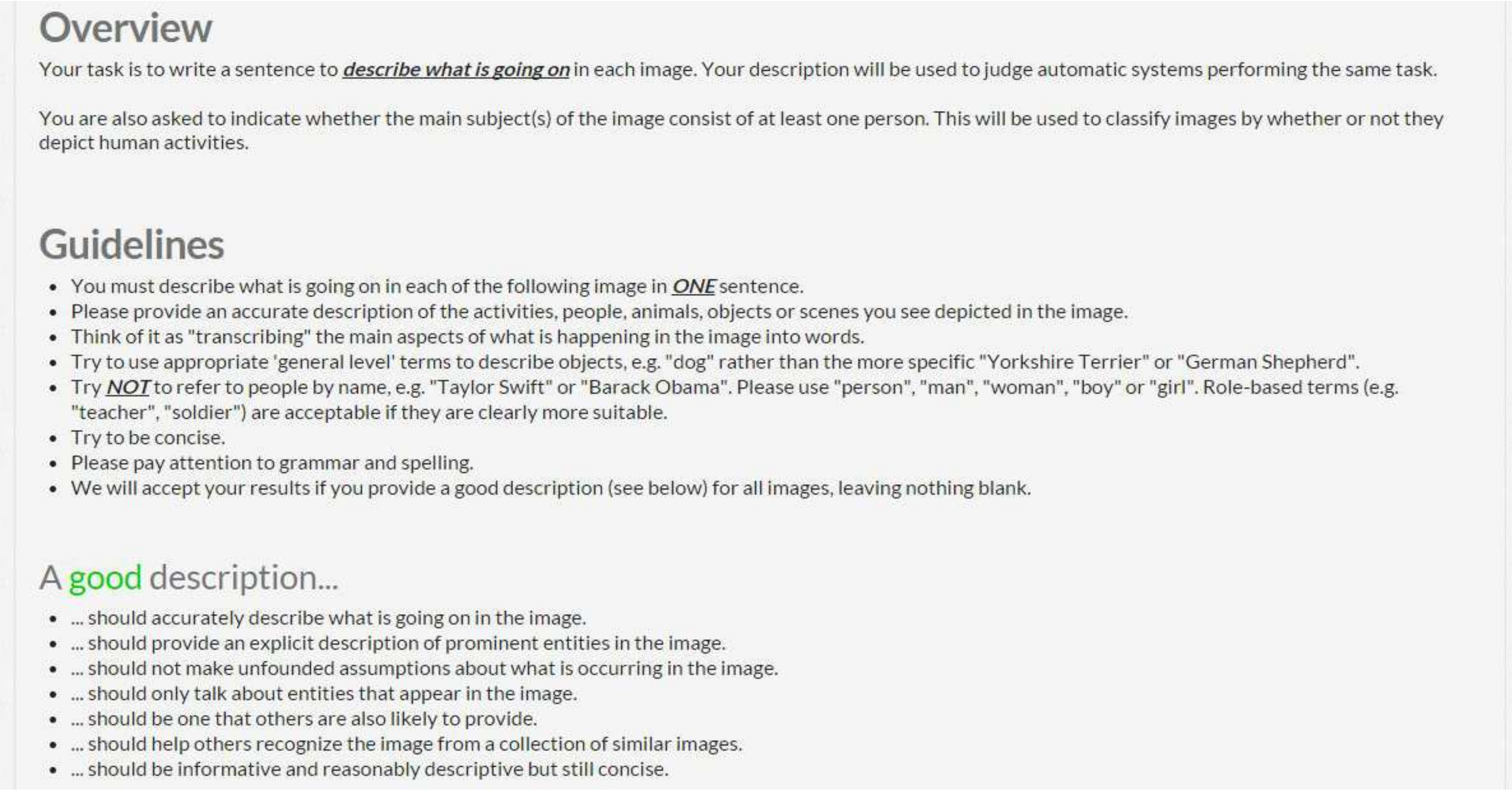
504 judgments (displaying first 200 only)

Number of boxes in the image: Please, use the confirm button to

Textual annotation

- Textual description annotations
  - Crowdflower: Minimum 5 sentences per image
  - Basic spell correction – using Aspell and going through them manually to confirm
  - Kept both American and British spellings (colour vs color) – as “real world challenge”
  - Development set contains 5 to 51 sentences per image
    - Mean: 9.492
    - Median: 8

Textual annotation



Overview

Your task is to write a sentence to ***describe what is going on*** in each image. Your description will be used to judge automatic systems performing the same task.

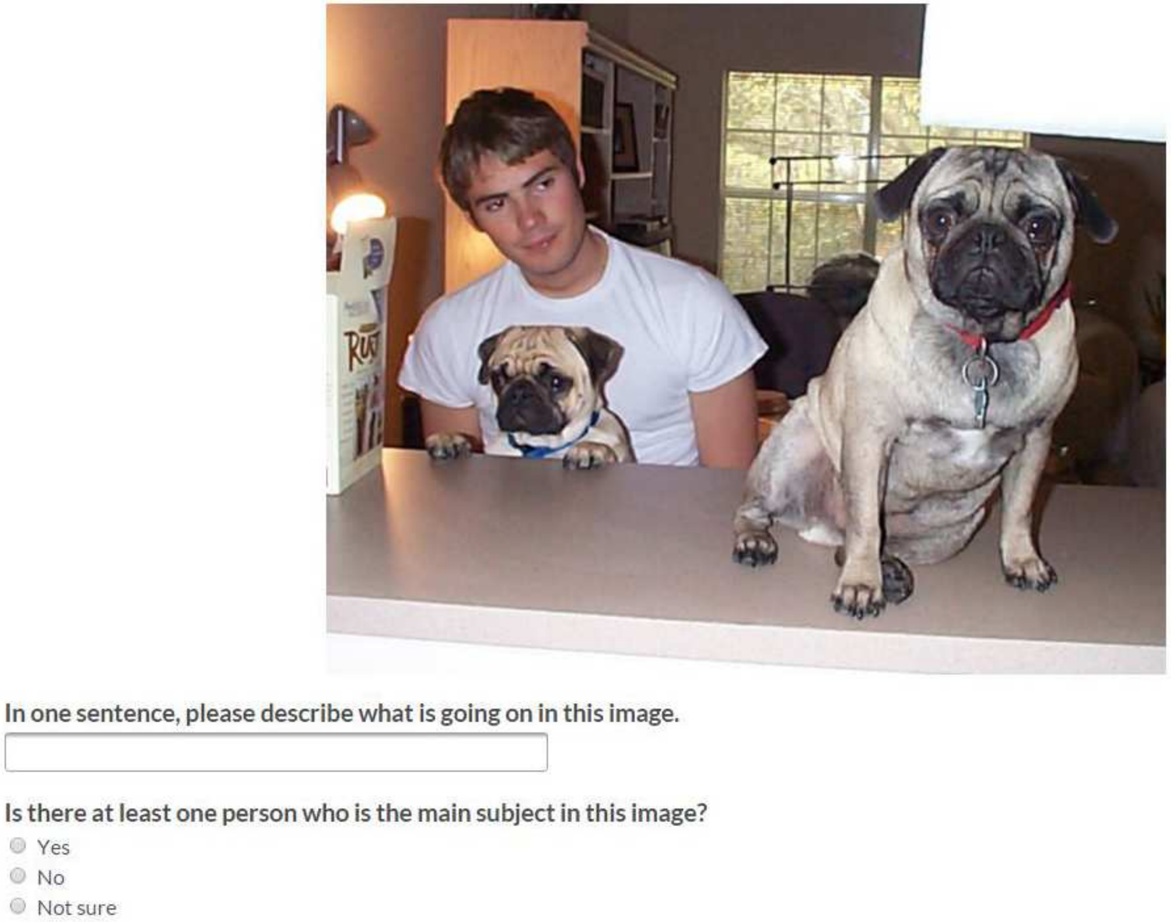
You are also asked to indicate whether the main subject(s) of the image consist of at least one person. This will be used to classify images by whether or not they depict human activities.

Guidelines

- You must describe what is going on in each of the following image in *ONE* sentence.
- Please provide an accurate description of the activities, people, animals, objects or scenes you see depicted in the image.
- Think of it as "transcribing" the main aspects of what is happening in the image into words.
- Try to use appropriate 'general level' terms to describe objects, e.g. "dog" rather than the more specific "Yorkshire Terrier" or "German Shepherd".
- Try *NOT* to refer to people by name, e.g. "Taylor Swift" or "Barack Obama". Please use "person", "man", "woman", "boy" or "girl". Role-based terms (e.g. "teacher", "soldier") are acceptable if they are clearly more suitable.
- Try to be concise.
- Please pay attention to grammar and spelling.
- We will accept your results if you provide a good description (see below) for all images, leaving nothing blank.

A good description...

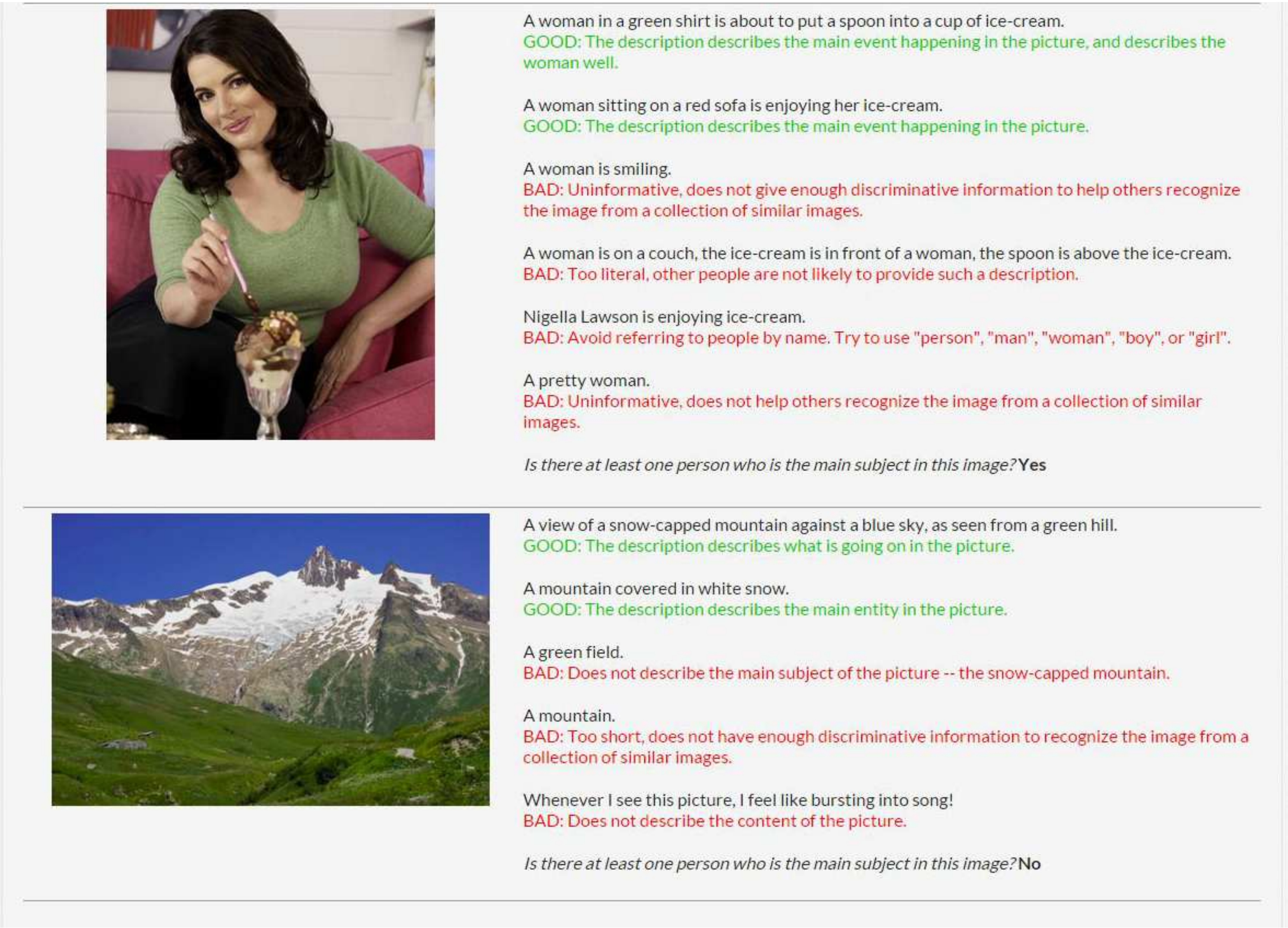
- ... should accurately describe what is going on in the image.
- ... should provide an explicit description of prominent entities in the image.
- ... should not make unfounded assumptions about what is occurring in the image.
- ... should only talk about entities that appear in the image.
- ... should be one that others are also likely to provide.
- ... should help others recognize the image from a collection of similar images.
- ... should be informative and reasonably descriptive but still concise.



In one sentence, please describe what is going on in this image.

Is there at least one person who is the main subject in this image?

○ Yes
○ No
○ Not sure



A woman in a green shirt is about to put a spoon into a cup of ice-cream.
GOOD: The description describes the main event happening in the picture, and describes the woman well.

A woman sitting on a red sofa is enjoying her ice-cream.
GOOD: The description describes the main event happening in the picture.

A woman is smiling.
BAD: Uninformative, does not give enough discriminative information to help others recognize the image from a collection of similar images.

A woman is on a couch, the ice-cream is in front of a woman, the spoon is above the ice-cream.
BAD: Too literal, other people are not likely to provide such a description.

Nigella Lawson is enjoying ice-cream.
BAD: Avoid referring to people by name. Try to use "person", "man", "woman", "boy", or "girl".

A pretty woman.
BAD: Uninformative, does not help others recognize the image from a collection of similar images.

Is there at least one person who is the main subject in this image? **Yes**

A view of a snow-capped mountain against a blue sky, as seen from a green hill.
GOOD: The description describes what is going on in the picture.

A mountain covered in white snow.
GOOD: The description describes the main entity in the picture.

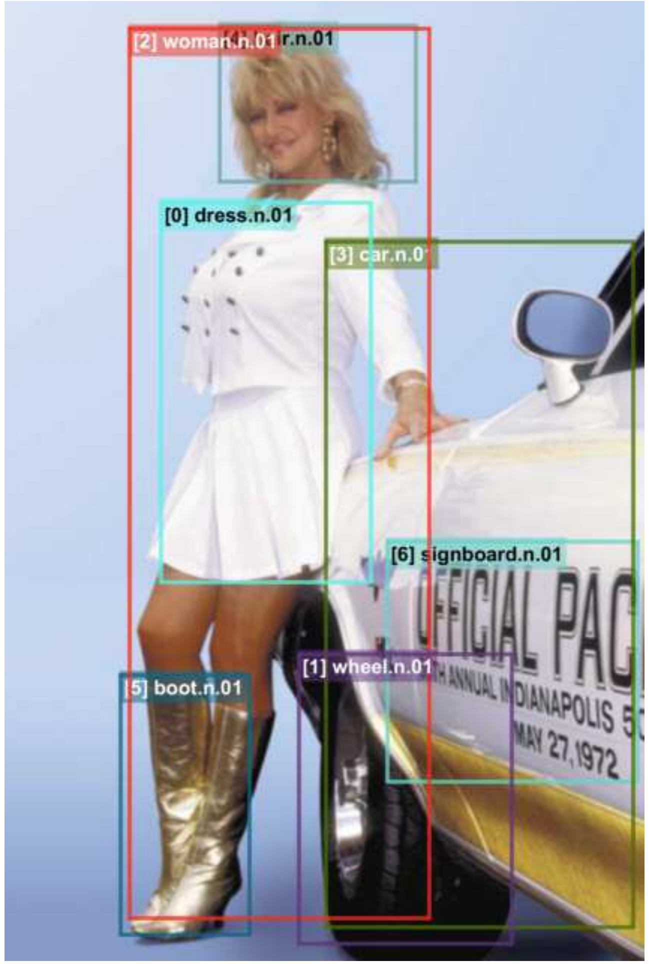
A green field.
BAD: Does not describe the main subject of the picture -- the snow-capped mountain.

A mountain.
BAD: Too short, does not have enough discriminative information to recognize the image from a collection of similar images.

Whenever I see this picture, I feel like bursting into song!
BAD: Does not describe the content of the picture.

Is there at least one person who is the main subject in this image? **No**

Textual Correspondence Annotation



A [[[woman|2]]] in a white [[[dress|0]]] and gold [[[boots|5]]] leaning on a [[[car|3]]] .

A [[[woman|2]]] poses along a [[[car|3]]] .

[[[Woman|2]]] dressed in white with gold [[[boots|5]]] poses next to a police [[[car|3]]]

A [[[woman|2]]] dressed in white leans against a white [[[car|3]]] .

A [[[woman|2]]] is leaning against a [[[car|3]]] .

A [[[woman|2]]] with gold [[[boots|5]]] leans against an Indy pace [[[car|3]]] .

A blonde [[[woman|2]]] wearing gold shiny [[[boots|5]]] , a white [[[top|0]]] and short white skirt is leaning on a [[[car|3]]] .

Difference to ILSVRC, MS COCO

- The training data is noisy
- Recognition/natural description generation based on images and text articles associated with images
- The test set is 500k

ImageCLEF 2015 - Features

- Image
  - CNN, GIST, Color Histograms, SIFT, C-SIFT, RGB-SIFT and OPPONENT-SIFT, BoW
- Text from the web pages where the images appear,
  - list of word-score pairs. The scores based on: 1) the term frequency (TF), 2) the document object model (DOM) attributes, and 3) the word distance to the image.
  - Triplets of word, search engine and rank, of how the images were found.
  - XML of the pages in which the images appeared.
  - The URLs of the images as referenced on the corresponding web pages (sometimes also relate to the content of the images).

Evaluation Protocol – sub task1

- Up to 100 localised Concepts with 100 Confidence based BBs per image allowed
- For image Localisation, intersection over Union (IoU) between GT and proposed localised concept
- The Confidence threshold was increased to provide a mean average precision (MAP) measure of performance

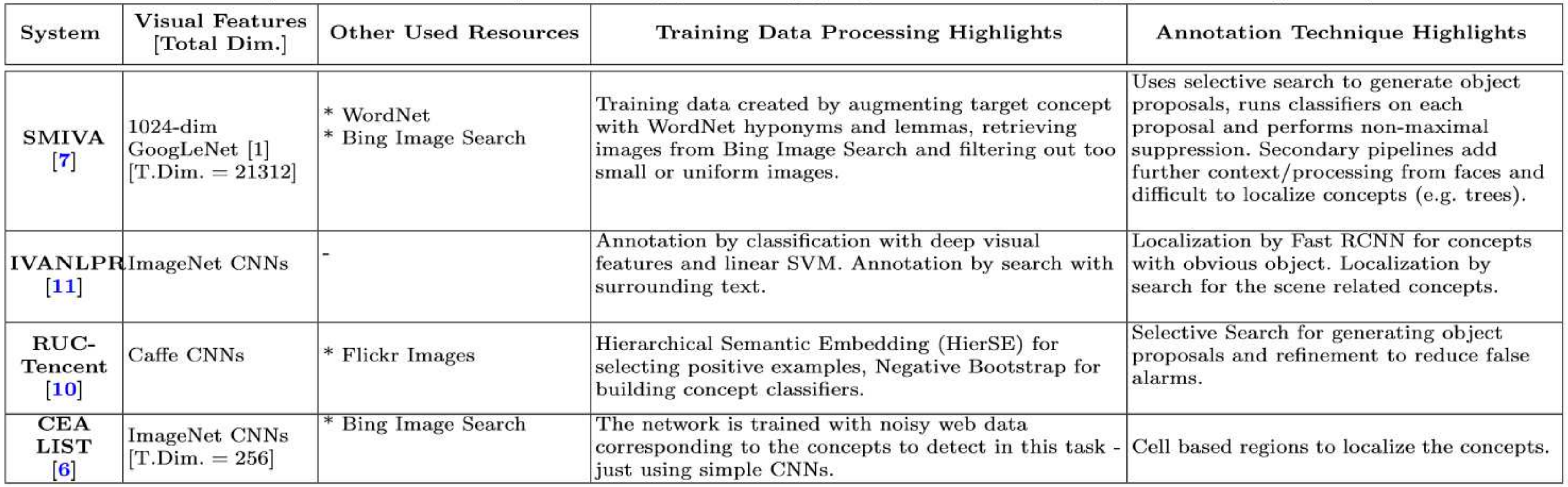
ImageCLEF sub task 2,3 Evaluation

- Meteor scoring
  - Identify content and function words in hypothesis and reference
  - For each matcher (exact, stem, synonym, paraphrase), count content & function words covered by the matcher
  - Compute harmonic mean of precision and recall
  - Penalise any fragmentation

Participation

- 14 groups, with 122 submitted runs, 11 working notes
- Including: China, France, Tunisia, Colombia, Japan, Romania,

Table 6: Key details of the best system for top performing groups that submitted a paper describing their system.



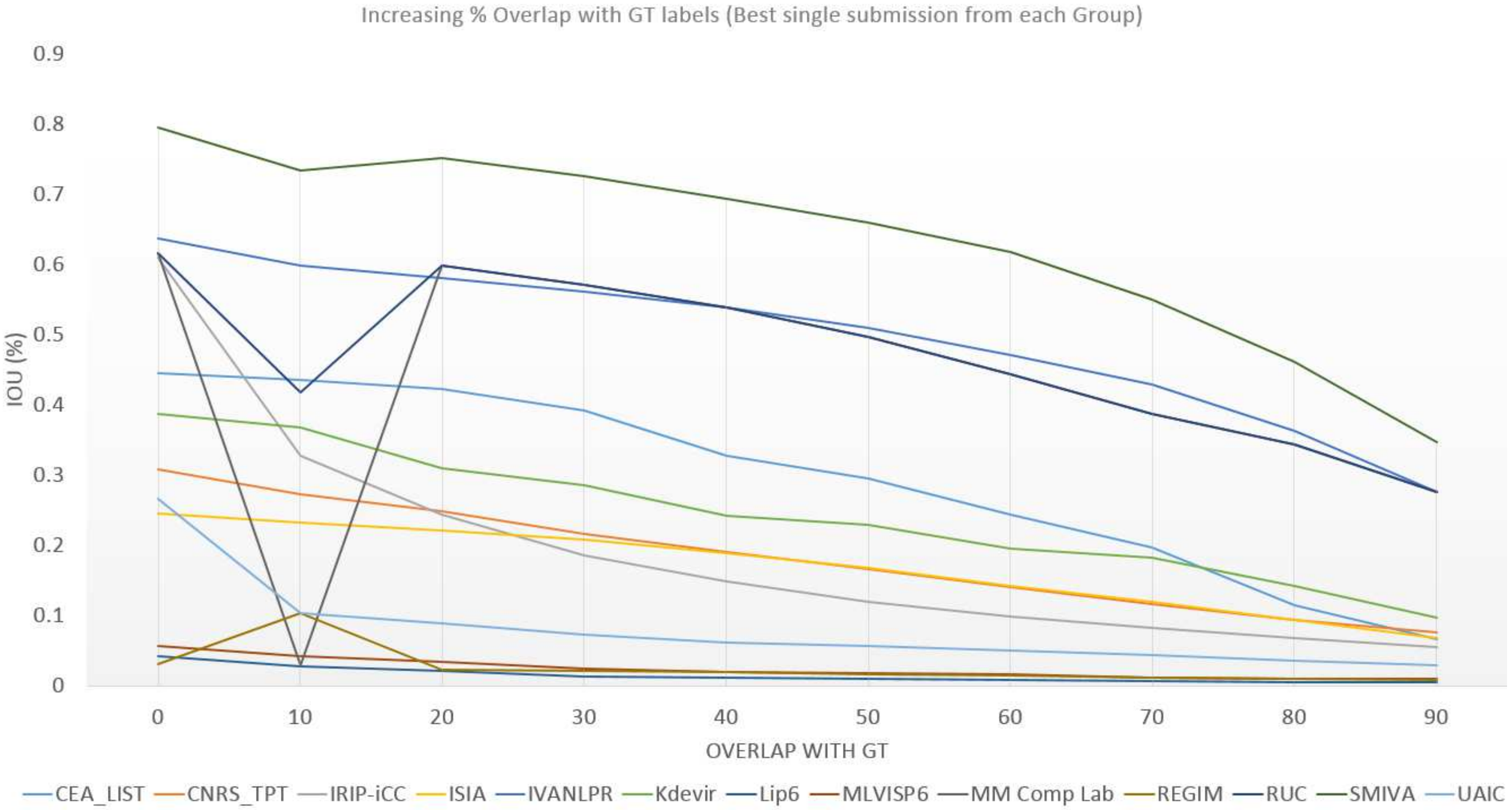
System	Visual Features [Total Dim.]	Other Used Resources	Training Data Processing Highlights	Annotation Technique Highlights
SMIVA [7]	1024-dim GoogLeNet [1] [T.Dim. = 21312]	* WordNet * Bing Image Search	Training data created by augmenting target concept with WordNet hyponyms and lemmas, retrieving images from Bing Image Search and filtering out too small or uniform images.	Uses selective search to generate object proposals, runs classifiers on each proposal and performs non-maximal suppression. Secondary pipelines add further context/processing from faces and difficult to localize concepts (e.g. trees).
IVANLPR [11]	ImageNet CNNs	-	Annotation by classification with deep visual features and linear SVM. Annotation by search with surrounding text.	Localization by Fast RCNN for concepts with obvious object. Localization by search for the scene related concepts.
RUC-Tencent [10]	Caffe CNNs	* Flickr Images	Hierarchical Semantic Embedding (HierSE) for selecting positive examples, Negative Bootstrap for building concept classifiers.	Selective Search for generating object proposals and refinement to reduce false alarms.
CEA LIST [6]	ImageNet CNNs [T.Dim. = 256]	* Bing Image Search	The network is trained with noisy web data corresponding to the concepts to detect in this task - just using simple CNNs.	Cell based regions to localize the concepts.

Results Sub task 1 – Overlap

Approach	0% Overlap	Approach	50% Overlap
/SMIVA/21.run	0.795403	/SMIVA/21.run	0.659507
/IVANLPR/gengxin	0.637086	/IVANLPR/gengxin	0.510028
/MM_Comp_Lab/ ruc_task1_svm_md3_nostar	0.616107	/MM_Comp_Lab/ ruc_task1_svm_md3_nostar	0.496057
/RUC/ruc_task1_svm_md3_nostar	0.616107	/RUC/ruc_task1_svm_md3_nostar	0.496057
IRIP-iCC/ sub9_vis_text_face_w6_prT_0.6	0.60918	/CEA_LIST/all_img.bb.run2	0.294559
/CEA_LIST/all_img.bb.run2	0.445121	/kdevir/run4.txt	0.228856
/kdevir/run4.txt	0.386693	/ISIA/run1_subtask1.txt	0.167836
/CNRS_TPT/TPT_RUN9.res	0.307308	/CNRS_TPT/TPT_RUN9.res	0.165816
/UAIC/output_rulare4_sortat_task1	0.265927	IRIP-iCC/ sub9_vis_text_face_w6_prT_0.6	0.119564
/ISIA/run1_subtask1.txt	0.245167	/UAIC/output_rulare4_sortat_task1	0.055917
/MLVISP6/run_mar1.txt	0.057422	/MLVISP6/run_mar1.txt	0.018619
/Lip6.fr/run_all_mar2.txt	0.041891	/REGIM/ regimvid_at_imageclef2015_task1_ 0.7.txt	0.016169
/REGIM/ regimvid_at_imageclef2015_task1_0. 7.txt	0.02054	/Lip6.fr/run_all_mar2.txt	0.010133

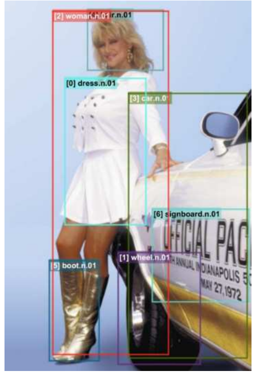
SMIVA - Social Media and Internet Vision Analytics Lab, Institute for Infocomm Research, Singapore

IVANLPR - National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences


Increasing % Overlap with GT labels (Best single submission from each Group)
IOU (%)
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
10
20
30
40
50
60
70
80
90
OVERLAP WITH GT
CEA_LIST
CNRS_TPT
IRIP-iCC
ISIA
IVANLPR
Kdevir
Lip6
MLVISP6
MM Comp Lab
REGIM
RUC
SMIVA
UAIC

Sub task 2,3

- Noisy track: Generate sentence for all 500k images
  - Meteor
- Clean track: Given bounding boxes (with concept labels) for 450 test images as input, generate sentences.
  - Meteor
  - Content selection F1 score
    - Did participants select the right bounding boxes to be described?



A [[[woman|2]]] with gold [[[boots|5]]] leans against an Indy pace [[[car|3]]] .

Gold Standard sentence g for image I

Predicted Sentence

$$P^{I} = \frac{1}{|G^{I}|} \sum_{g=0}^{|G^{I}|} \frac{|G_{g}^{I} \cap S|}{|S|} \quad R^{I} = \frac{1}{|G^{I}|} \sum_{g=0}^{|G^{I}|} \frac{|G_{g}^{I} \cap S|}{|G_{g}^{I}|} \quad F^{I} = 2 \times P^{I} \times R^{I} / P^{I} + R$$

ImageCLEF 2015: Sentence Generation

Subtask 2 Clean Track: Content Selection scores (Average F1 scores)

	MEAN PRECISION	MEAN RECALL	MEAN F1
Human	*0.7690 ± 0.1090*	*0.7690 ± 0.1090*	*0.7445 ± 0.1174*
RUC	0.7015 ± 0.3095	0.4496 ± 0.2488	0.5147 ± 0.2390
UAIC	0.5095 ± 0.1938	0.5547 ± 0.2415	0.5030 ± 0.1775
Baseline	*0.1983 ± 0.2003*	*0.1817 ± 0.2227*	*0.1800 ± 0.1973*

RUC - Multimedia Computing Lab,
School of Information,
Renmin University of China

UAIC - Faculty of Computer Science,
“Alexandru Ioan Cuza” University,
Romania

Human upper-bound: One gold standard against others (for same image), repeat and average
Baseline: Select random 3 bounding boxes from gold input, connect concept terms with random prepositions/ conjunctions followed by an optional article “the”

ImageCLEF 2015: Sentence Generation

Subtask 2 Clean Track: METEOR scores

	MEAN ± STD	MEDIAN	MIN	MAX
Human	*0.4786 ± 0.1706*	*0.4420*	*0.2353*	1.0000
RUC	0.2393 ± 0.0865	0.2278	0.0598	0.5745
UAIC	0.2097 ± 0.0660	0.2085	0.0290	0.7246
Baseline	*0.0977 ± 0.0467*	*0.0888*	*0.0237*	*0.3102*

RUC - Multimedia Computing Lab,
School of Information,
Renmin University of China

UAIC - Faculty of Computer Science,
"Alexandru Ioan Cuza" University,
Romania

Human upper-bound: One gold standard against others (for same image), repeat and average
Baseline: Select random 3 bounding boxes from gold input, connect concept terms with random prepositions/ conjunctions followed by an optional article "the"

ImageCLEF 2015: Sentence Generation

Subtask 2 Noisy Track: METEOR scores

	MEAN ± STD	MEDIAN	MIN	MAX
Human	*0.3385 ± 0.1556*	*0.3355*	*0.0000*	*1.0000*
RUC	0.1875 ± 0.0831	0.1744	0.0201	0.5696
ISIA	0.1687 ± 0.0852	0.1529	0.0387	1.0000
Baseline (CNN+LSTM)	*0.1490 ± 0.0741*	*0.1364*	*0.0189*	*0.5696*
MindLab	0.1403 ± 0.0564	0.1342	0.0256	0.3745
UAIC	0.0813 ± 0.0513	0.0769	0.0142	0.3234

RUC - Multimedia Computing Lab, School of Information, Renmin University of China

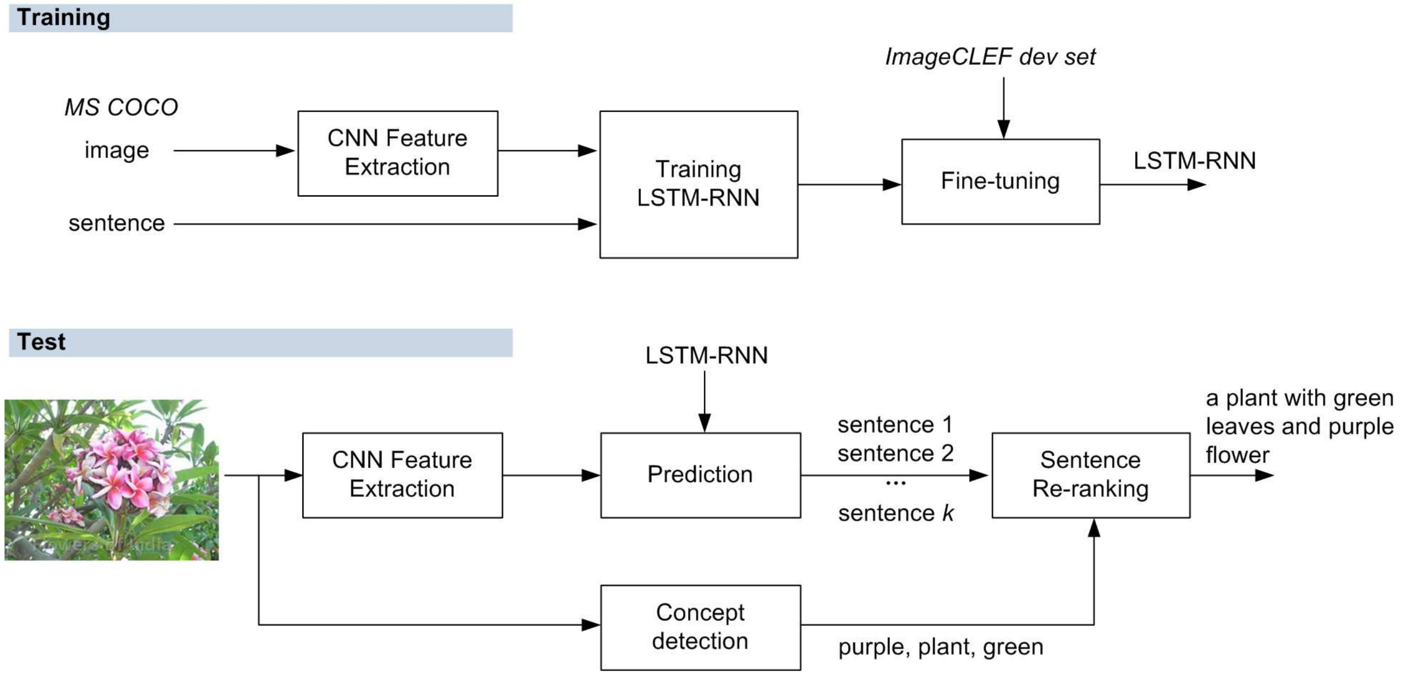
ISIA - Visual Information Processing and Learning Institute of Computing Technology, Chinese Academy of Sciences

MindLab - INAOE in Mexico and UNAL in Colombia

UAIC - Faculty of Computer Science, "Alexandru Ioan Cuza" University, Romania

Human upper-bound: One gold standard against others (for same image), repeat and average
Baseline: Stanford NeuralTalk (untuned)

RUC Image Sentence Generation



Training

MS COCO

image

sentence

CNN Feature Extraction

Training LSTM-RNN

ImageCLEF dev set

Fine-tuning

LSTM-RNN

Test

CNN Feature Extraction

LSTM-RNN

Prediction

sentence 1
sentence 2
…
sentence *k*

Sentence Re-ranking

a plant with green leaves and purple flower

Concept detection

purple, plant, green

Ingredient

+ **Google's LSTM-RNN** (from NeuralTalk) for sentence encoding and image decoding

+ **VGGNet CNN** for image representation

+ **Sentence re-ranking** by concept detection

Summary

- New data and challenge for image caption generation
- Different than Flickr30k or MS COCO
- Also addressed towards NLP community
- The data size and category labels will grow in new editions