Overview of the ImageCLEF 2016 Scalable Concept Image Annotation Task

Andrew Gilbert, Luca Piras, Josiah Wang*, Fei Yan*, Arnau Ramisa*, Emmanuel Dellandrea*, Robert Gaizauskas*, Mauricio Villegas & Krystian Mikolajczyk*

* chist-era The Visual Sense (ViSen) consortium

Motivation and Aim

- Motivations:
  - Users struggle with the ever-increasing quantity of data available to them
  - Large amounts of images can be cheaply found and gathered from the Internet
  - Web pages can provide both images and text a more valuable mixed modality data

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

ImageCLEF 2016: Tasks Overview

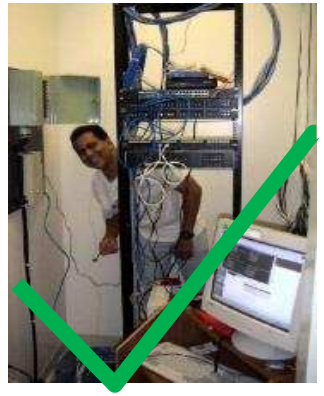
- Subtask 1: Image annotation and localization
- Subtask 2: Natural language caption generation
- Subtask 3: Content selection
- Teaser 1: Text illustration (New for 2016)

- Details of each task later

ImageCLEF 2016: Dataset

- Single noisy dataset of 510K webpages, images + text
- Subtasks 1 & 2
  - Test dataset $\subset$ 510K training dataset
- Subtask 3
  - Training & test datasets taken from 510K dataset
- Teaser 1
  - 510K dataset split into 310K for training and 200K for testing

Training, development and test data



- 251 concepts from airplane to bottle to face & arm
  - Formed from looking at word co-occurrence in 34M webpages of all English dictionary words
- Training/test set of 510K images, >20 images per concept
  - CNN trained to identify “interesting images” especially for natural sentence generation
- The development set contained 2000 images.
- Labelled test of 3070 images (subtask1, 2), 450 (subtask 3), both within the 510K
- In total: crowd sourced annotation of 5500 images
  - BBs of single instances or grouped instances
  - Annotations are not exhaustive

Differences to ILSVRC, MS COCO

- The training data is ”real”, noisy and mixed modality
- Recognition/natural description generation is based on images and text articles associated with images
- The test set is 510k (Subtasks 1 & 2)

Image Features

- Pre-computed image features
  - Color Histograms
  - GIST
  - SIFT, C-SIFT, RGB-SIFT and OPPONENT-SIFT → BoW (Pyramid)
  - **CNN** (AlexNet fc7)

Text Features

- Pre-computed text features (from webpages)
  - Triplets of<word, search engine, rank>, of how each image was found.
  - Image URL on webpage (might relate to image content)
  - The webpage (converted to valid XML)
  - <word, score> of nearby text. Scores based on:
    - term frequency (TF)
    - DOM attributes (title, alt, etc.)
    - Spaital distance to image on web page

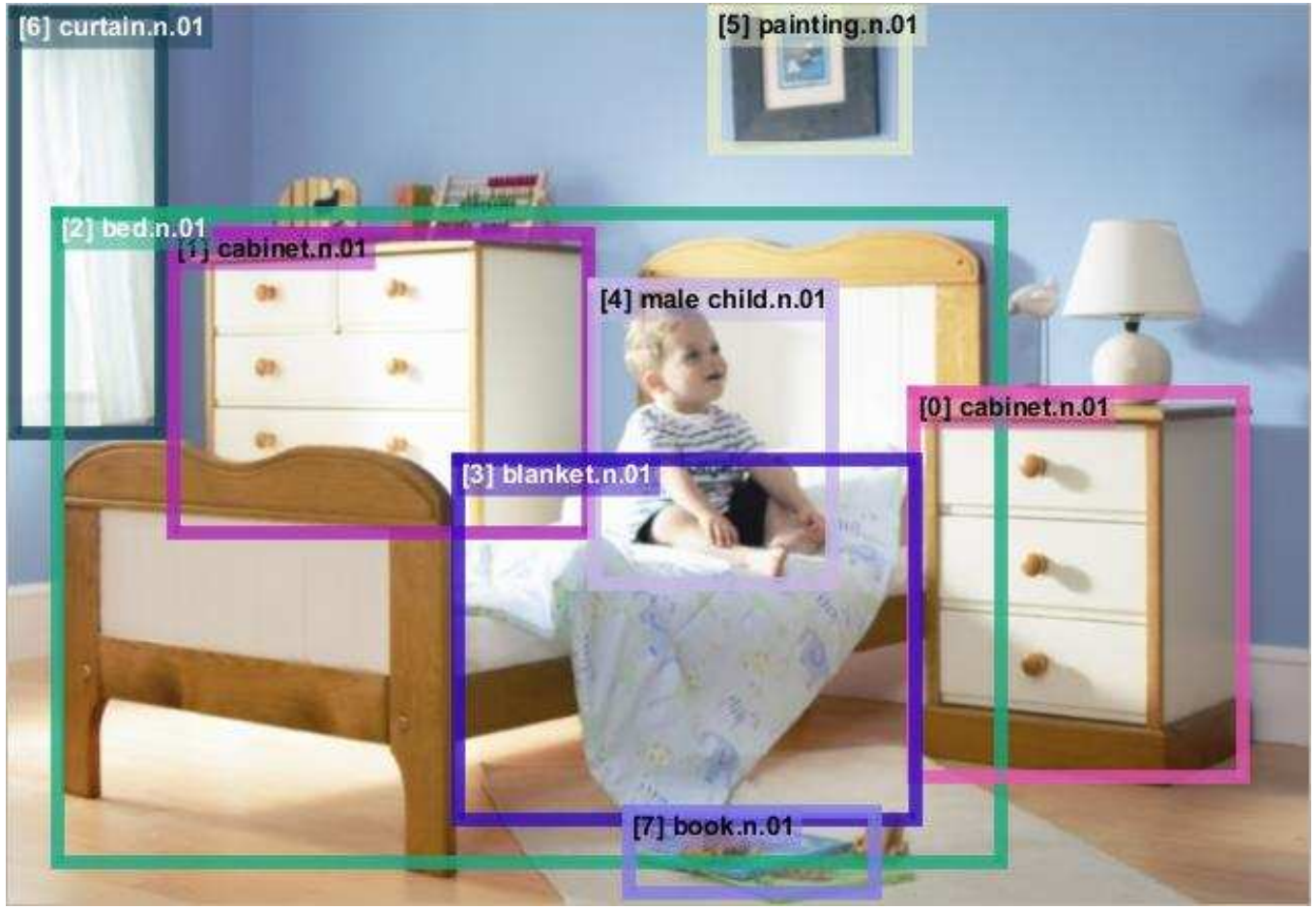
Participation

- 7 groups, with 50 submitted runs, 7 working notes

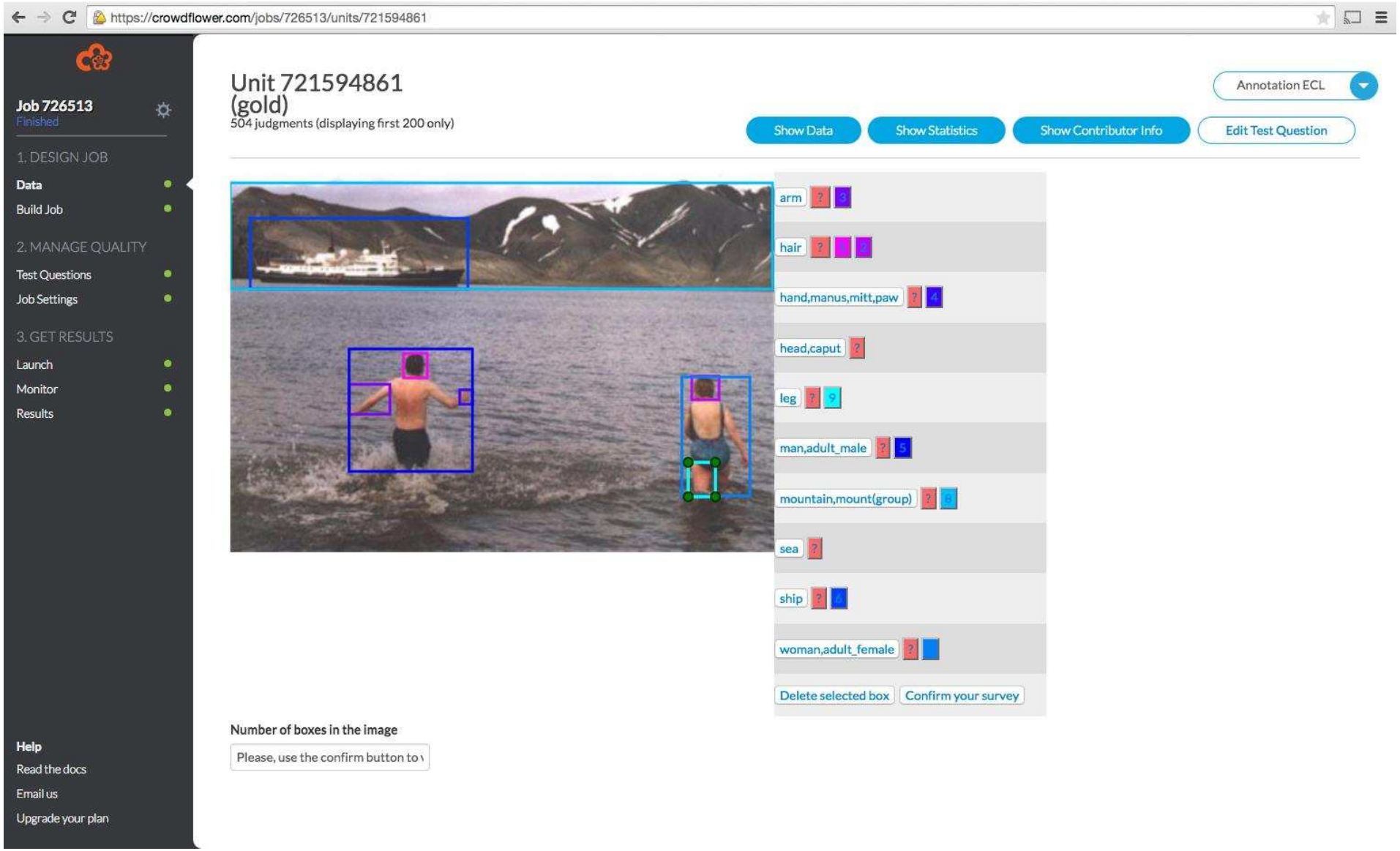
- CEA LIST: France
- CNRS TPT: France – Presentation & Poster
- DUTh: Greece
- ICTisia: China - Presentation
- INAOE: Mexico - Presentation
- MRIM-LIG: France - Presentation
- UAIC: Romania

Subtask 1: Image Annotation + Localization

- Annotate & localize 251 concepts in 510K images



Subtask 1: Bounding box annotation



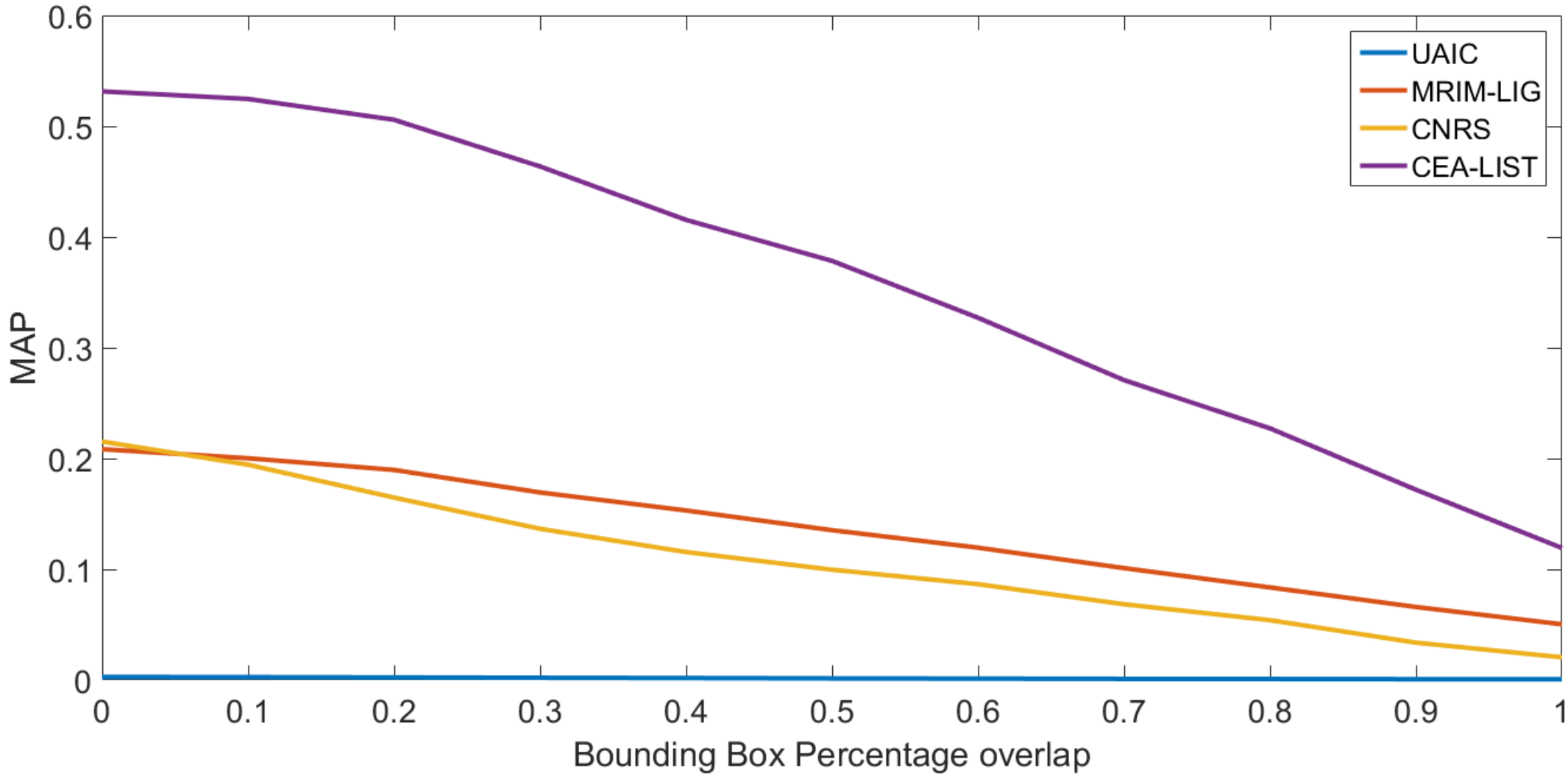
Subtask 1: Evaluation

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

Subtask 1: Results - Overlap

Table 1: Subtask 1 results.

Group	0% Overlap	50% Overlap
CEA LIST	0.54	0.378
MRIM-LIG	0.21	0.14
CNRS	0.25	0.11
UAIC	0.003	0.002


0.6
0.5
0.4
0.3
0.2
0.1
0
MAP
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Bounding Box Percentage overlap
UAIC
MRIM-LIG
CNRS
CEA-LIST

Subtask 1: Per Concept

Concept	Ave MAP across all Groups		% of Occurrence
	0.0 BB Overlap	0.5 BB Overlap	in test images
Ship	0.61	0.57	28.0%
Car	0.62	0.55	25.3%
Airplane	0.60	0.55	3.2%
Hair	0.74	0.52	93.0%
Park	0.41	0.52	13.9%
Floor	0.41	0.51	13.4%
Boot	0.43	0.59	4.2%
Sea	0.45	0.49	8.8%
Street	0.54	0.47	18.0%
Face	0.75	0.47	95.7%
Street	0.64	0.45	59.9%

No method managed to localise 38 concepts, these include the concepts: nut, mushroom, banana, ribbon, planet, milk, orange fruit and strawberry.

Subtask 1: Best Systems

System	Visual Features	Other Used Resources	Training Data Processing Highlights	Annotation Technique Highlights
CEA LIST [2]	16-layer CNN 50-layer ResNet	* Bing Image Search	They collected a set of roughly 251,000 images (1,000 images per concept) from the Bing Images search engine. For each concept they used its name and its synonyms (if present) to query the search engine. They used 90% of the dataset for training and 10% for validation.	They used EdgeBoxes, a generic objectness object detector, extracting a maximum of 100 regions per image then feeding each one to the CNN models. The concept that had the highest probability among the 251 concepts it has been kept.
MRIM-LIG [15]	152-layer ResNet	* Bing Image Search	Two-step learning process using two validation sets. First set of training images,learn the last layer of CNN. Retrain using 200 additional training images defined by the authors according to the low quality recognition concepts	An apriori set of bounding boxes which are expected to contain a single concept each is defined. Each of these boxes have been used as an input image on which the CNN has been applied to detect objects. Localization of parts of faces is achieved through the Viola and Jones approach and facial landmarks detection.
CNRS [18]	VGG deep network	* Google Image Search	2,000 images of the dev set have been used in order to enrich the labels of all the training set transferring the knowledge about the co-occurrence of some labels.	For each concept it has been trained "one-versus-all" SVM classifier.

Subtask 2: Caption Generation

- Generate image description (510k images)

A boy sitting on a bed in a bedroom.

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**

Subtask 2: Evaluation

- **Meteor** evaluation metric
- Adapted from Machine Translation

Subtask 2: Results

	MEAN ± STD	MEDIAN	MIN	MAX
Human	*0.3385 ± 0.1556*	*0.3355*	*0.0000*	*1.0000*
RUC (2015)	0.1875 ± 0.0831	0.1744	0.0201	0.5696
ICTisia	**0.1837 ± 0.0847**	**0.1711**	**0.0180**	**0.5934**
Baseline (CNN+LSTM)	*0.1490 ± 0.0741*	*0.1364*	*0.0189*	*0.5696*
UAIC	**0.0934 ± 0.0249**	**0.0915**	**0.0194**	**0.2514**
UAIC (2015)	0.0813 ± 0.0513	0.0769	0.0142	0.3234

ICTisia - Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, China

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

ImageCLEF 2015:

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

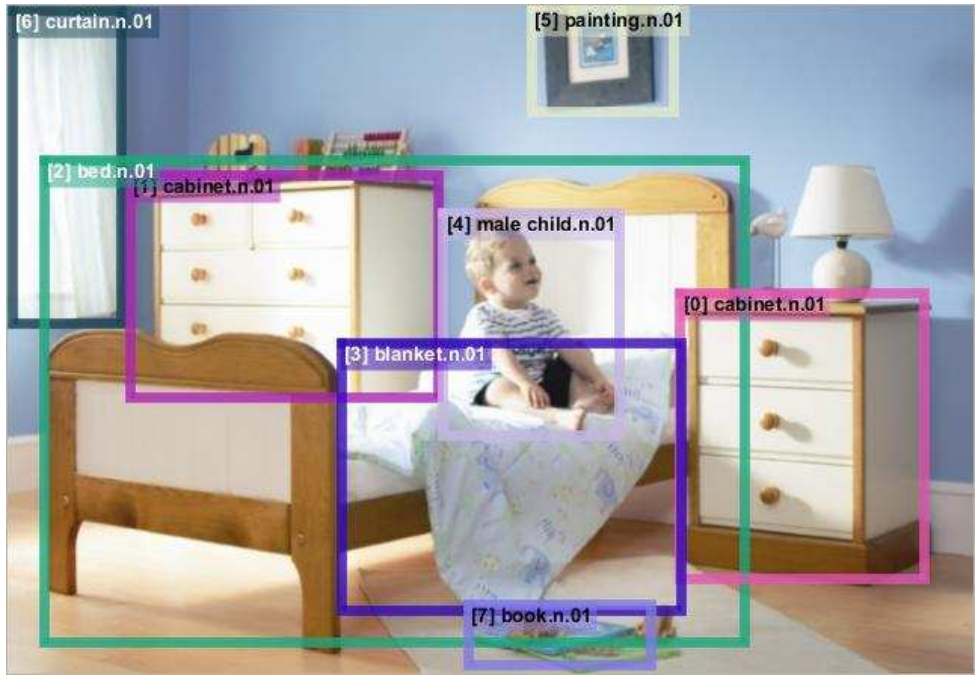
Human upper-bound: One gold standard against others (for same image), repeat and average
Baseline: Stanford NeuralTalk (untuned)

Subtask 2: Systems Overview

System	Visual Representation	Textual Representation	Other Used Resources	Summary
ICTisia [29]	VGGNet (FC7)	LSTM	* MSCOCO * Flickr8K/Flickr30K * Manually selected image-caption pairs, captions generated from training set	CNN-LSTM caption generator, but fine-tuning *both* CNN and LSTM parameters. Also fine-tune on different datasets.
UAIC [3]	TensorFlow CNN (architecture unknown)	Text labels	* Face recognition module * WordNet * DuckDuckGo	Concept detection using textual features and visual feature (subtask 1), and generate descriptions using templates (with backoff).

Subtask 3: Content Selection

- Given labelled bounding box input (450 test images), select the instances most likely to be mentioned by humans in a description



[4] male_child.n.01
[3] blanket.n.01
[2] bed.n.01

Subtask 3: Evaluation

- Content Selection score (F-score)

bbox instances referenced in gold standard

bbox instances referenced in generated sentence

$$P^I = \frac{1}{M} \sum_{m=1}^{M} \frac{|\, G_m^I \cap S^I \,|}{|S^I|} \qquad R^I = \frac{1}{M} \sum_{m=1}^{M} \frac{|\, G_m^I \cap S^I \,|}{|G_m^I|} \qquad F^I = 2 \times \frac{P^I \times R^I}{P^I + R^I}$$

 # gold standard descriptions for image *I*

- Final score: Average over all test images

Subtask 3: Results

	MEAN F1	MEAN PRECISION	MEAN RECALL
Human	*0.7445 ± 0.1174*	*0.7690 ± 0.1090*	*0.7690 ± 0.1090*
DUTh	**0.5459 ± 0.1533**	**0.4451 ± 0.1695**	**0.7914 ± 0.1960**
RUC (2015)	0.5147 ± 0.2390	0.7015 ± 0.3095	0.4496 ± 0.2488
UAIC (2015)	0.5030 ± 0.1775	0.5095 ± 0.1938	0.5547 ± 0.2415
UAIC	**0.4982 ± 0.1782**	**0.4597 ± 0.1553**	**0.5951 ± 0.2592**
Baseline	*0.1800 ± 0.1973*	*0.1983 ± 0.2003*	*0.1817 ± 0.2227*

DUTh - Democritus University of Thrace, Greece

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

ImageCLEF 2015:
RUC - Multimedia Computing Lab, School of Information, Renmin University of China

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"

Subtask 3: Systems Overview

System	Representation	Content Selection Algorithm	Summary
DUTh [1]	* Bounding box (Position, size) * Local descriptors (Canny, Harris, BRISK, SURF, FAST) * Entropy	Nonlinear, binary SVM classifier (RBF, Polynomial kernels)	SVM classifier to classify whether a bounding box is important/not important, using combinations of local features.
UAIC [3]	Text labels	Selection by generating descriptions (subtask 2).	Bounding box selection by selecting up to three tuples (concept1, verb, concept2).

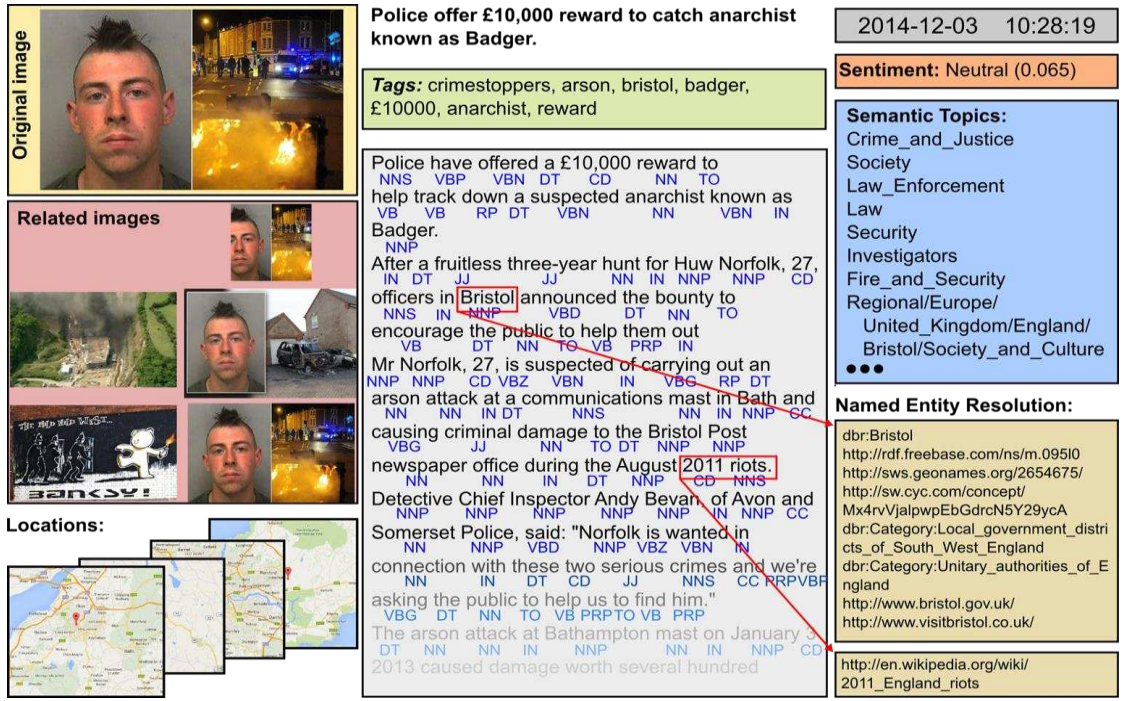
Teaser task: Text Illustration

- Given a piece of text, select the most relevant image to illustrate it (out of 200k test images)

Weather: winter proper is finally on its way with snow and ice. Winter proper is finally on its way with snow, frost, ice and fog forecast for next week. Wind and rain have blighted the country for weeks as thousands of homes have flooded and wet and windy conditions and brought chaos to our roads and transport systems. There is finally an end to the downpour in sight, but the change will bring the wintry conditions which have been held off by the storms, forecasters predict. Until then the country is at risk of further flooding as "slow responding rivers" are hit by further the heavy rain, although the worst is over. A pensioner today became the …

Teaser task: Text Illustration

- Based on the BreakingNews dataset
  - Arnau Ramisa et al. (2016) BreakingNews: Article Annotation by Image and Text Processing [Arxiv 1603.07141]



~100K news articles

- News article text
- Captions
- Reader comments
- Original image
- Related images from Google
- Tags
- GPS coordinates
- Linguistic features (PoS tags, named entities, semantic topics list etc.)

Teaser task: Text Illustration

- 510K dataset contains ~10K webpage-image pairs from BreakingNews.
- ‘Webpage’ generated from news text with a template
- Split into 310K for training and 200K for testing (~10K in test)
- Separate development set (~3K) provided
- Text extracted from 180K test webpages as test input

Teaser task: Evaluation

- Evaluated using Recall @ kth rank position (**R@k**) of the ground truth image

- For each 180K test documents, participants provided top 100 ranked images
  - Several values of *k* tested

Teaser task: Results

	Test set	R@1	R@10	R@50	R@100
Random chance		*0.00*	*0.01*	*0.03*	*0.05*
CEA LIST	10K (News)	0.02	0.11	0.46	0.80
	180K (Full)	0.18	1.05	3.00	4.51
INAOE	10K (News)	37.05	78.06	79.74	79.77
	180K (Full)	28.75	75.48	86.79	87.59

CEA LIST - CEA, LIST, Laboratory of Vision and Content Engineering, France
INAOE - Instituto Nacional de Astrofisica, Optica y Electronica (INAOE), Mexico

Results not quite directly comparable because INAOE used the test webpages at test time while CEA LIST did not

Teaser task: Systems Overview

System	Visual Representation	Textual Representation	Other Used Resources	Summary
CEA LIST [2]	VGGNet (FC7)	word2vec (TF-IDF weighted average)	* WordNet * Flickr Groups * NLTK Pos Tagger	Two methods: (i) Semantic signature - fixed sized vector, each element corresponding to a semantic concept. Inverse indexing for retrieval. (ii) Projection onto common, bimodal latent space via kCCA.
INAOE [14]	–	* TF-IDF weighted bag of words * word2vec (simple average)	–	IR queries using bag-of-words or word2vec.

Summary

- Reasonable improvements somewhat driven by newer CNN models
- New data and challenge for image caption generation
- Different than Flickr30k, ILSVRC or MS COCO
- Also addressed towards NLP community
- The data size and category labels will grow in new editions, & improve annotation recall levels