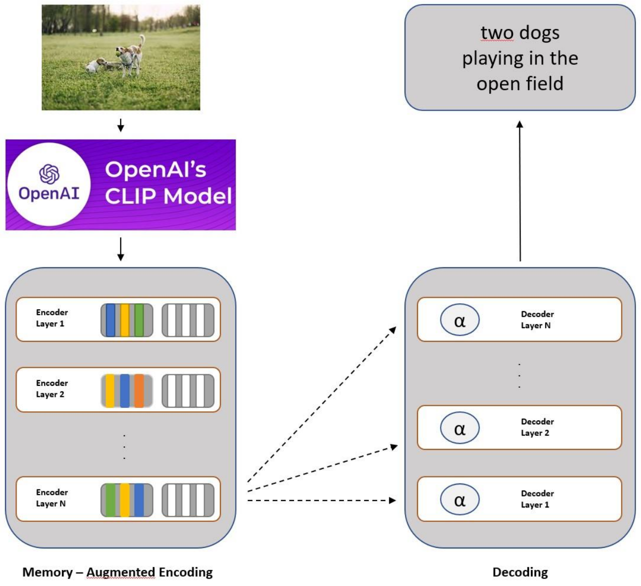
Course of:

Artificial Intelligence for Automotive

Prof.ssa Rita Cucchiara

Prof. Lorenzo Baraldi

Image Captioning in the Automotive Domain



Presentation by:

Federico Cocchi

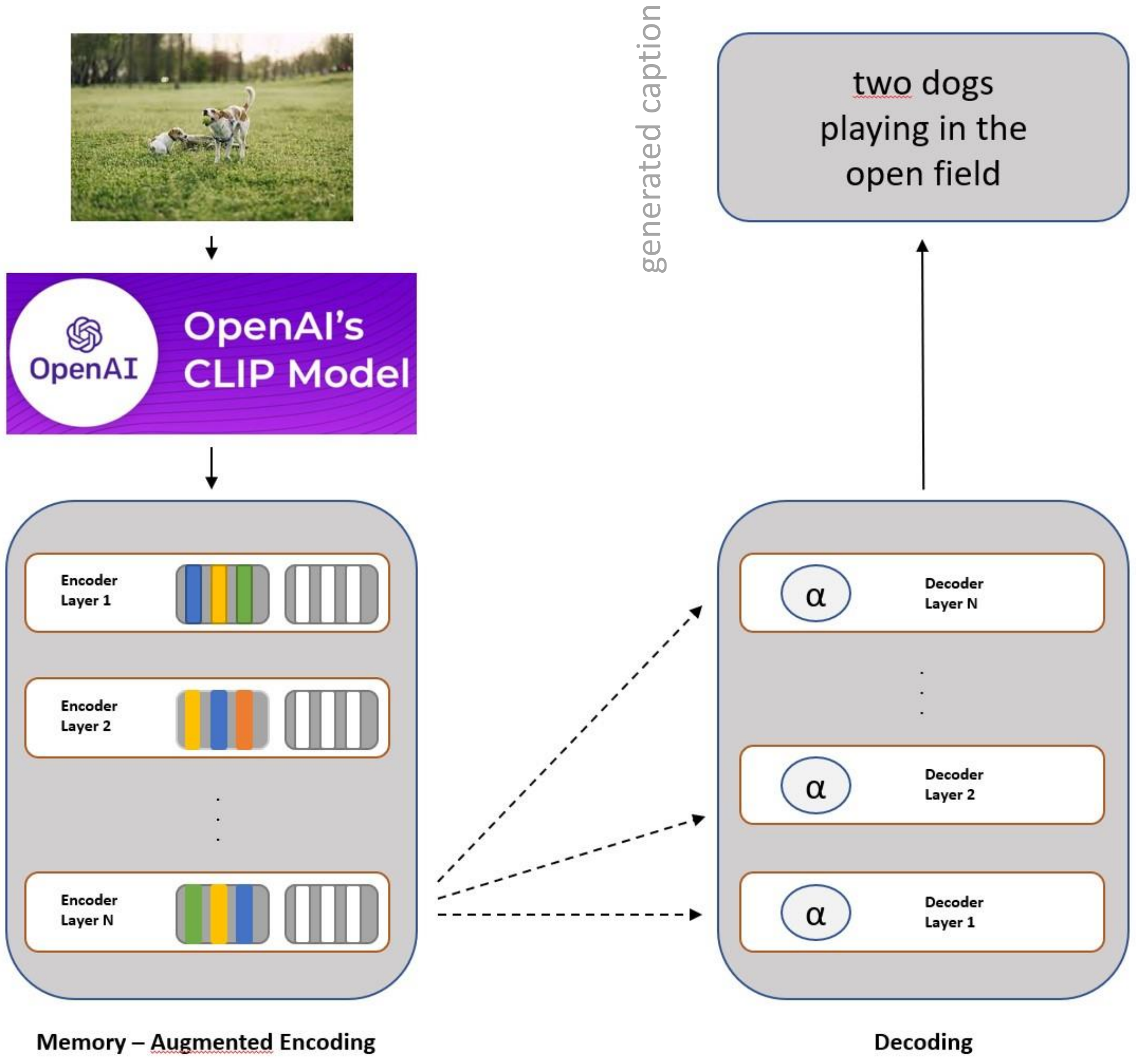
Paula Klinke

25.02.2022

1. Image Captioning description
2. Dataset
3. Architecture
4. Training
5. Results
6. Demo and qualitative results

Structure

Image Captioning



Starting from an **image**, the system generate a description of it using the **language**

Methods for Image Captioning:

- RNN
- CNN
- **Transformer**

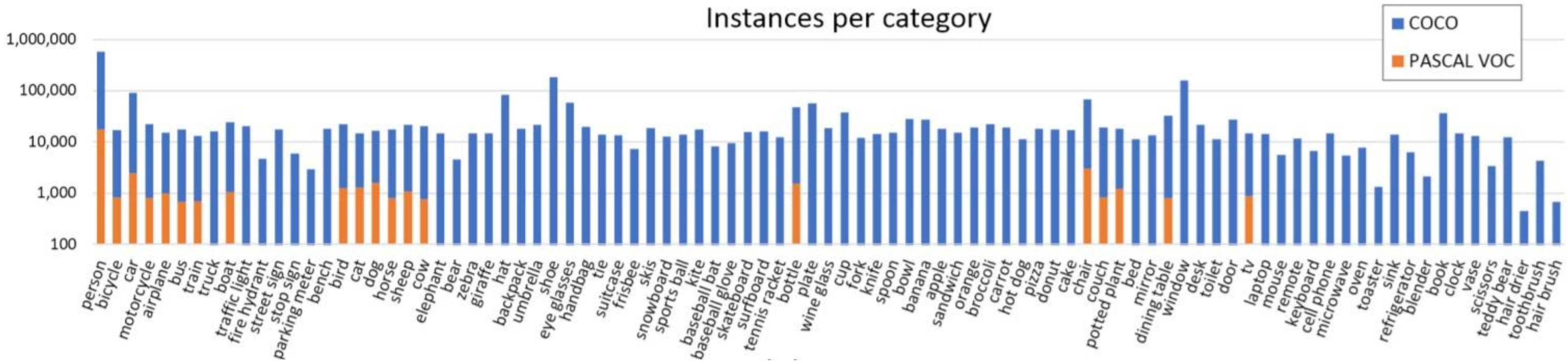
Dataset on automotive

Dataset in literature for Image captioning

- CC-3M
- WIT
- YFCC100M
- COCO

Types of data necessary for our project

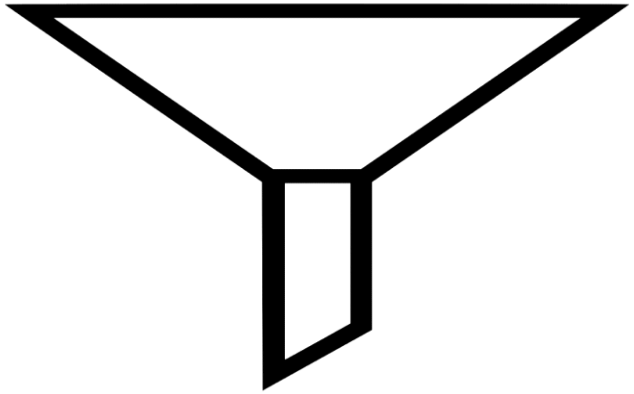
```
img_name.jpg
img_name.tags.json
img_name.txt
```



COCO distribution of classes

Dataset

Conceptual Captions



auto, car, automotive, street, road, parking, highway, semaphore, pedestrian, taxi, vehicle

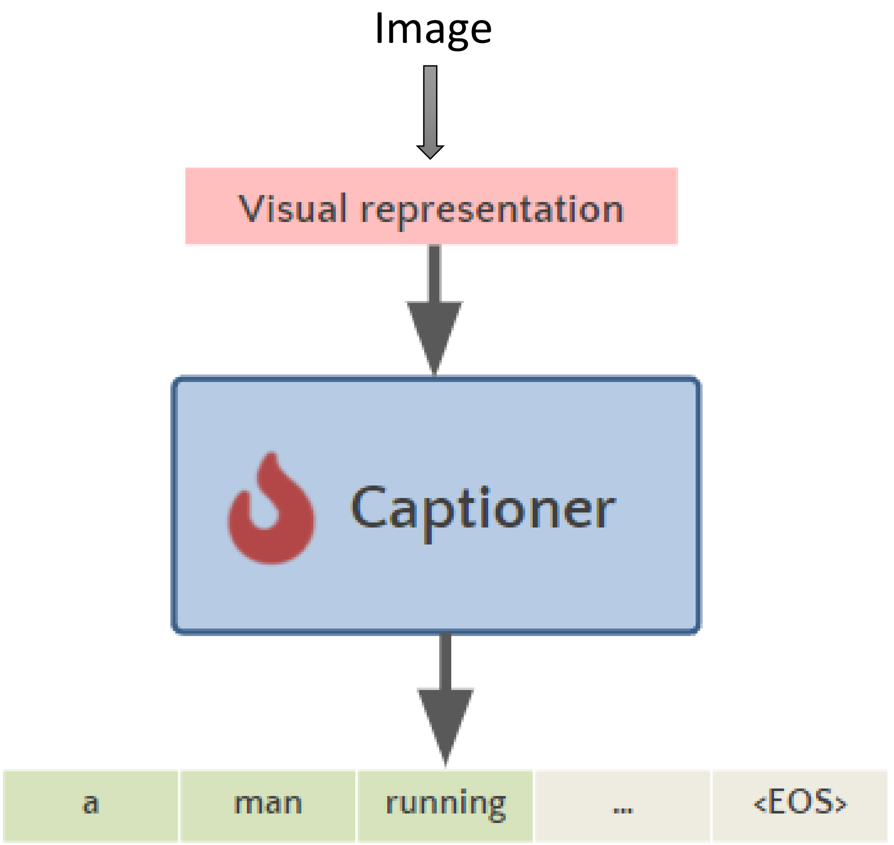
CC_automotive
155K data

aerial shot over a busy highway

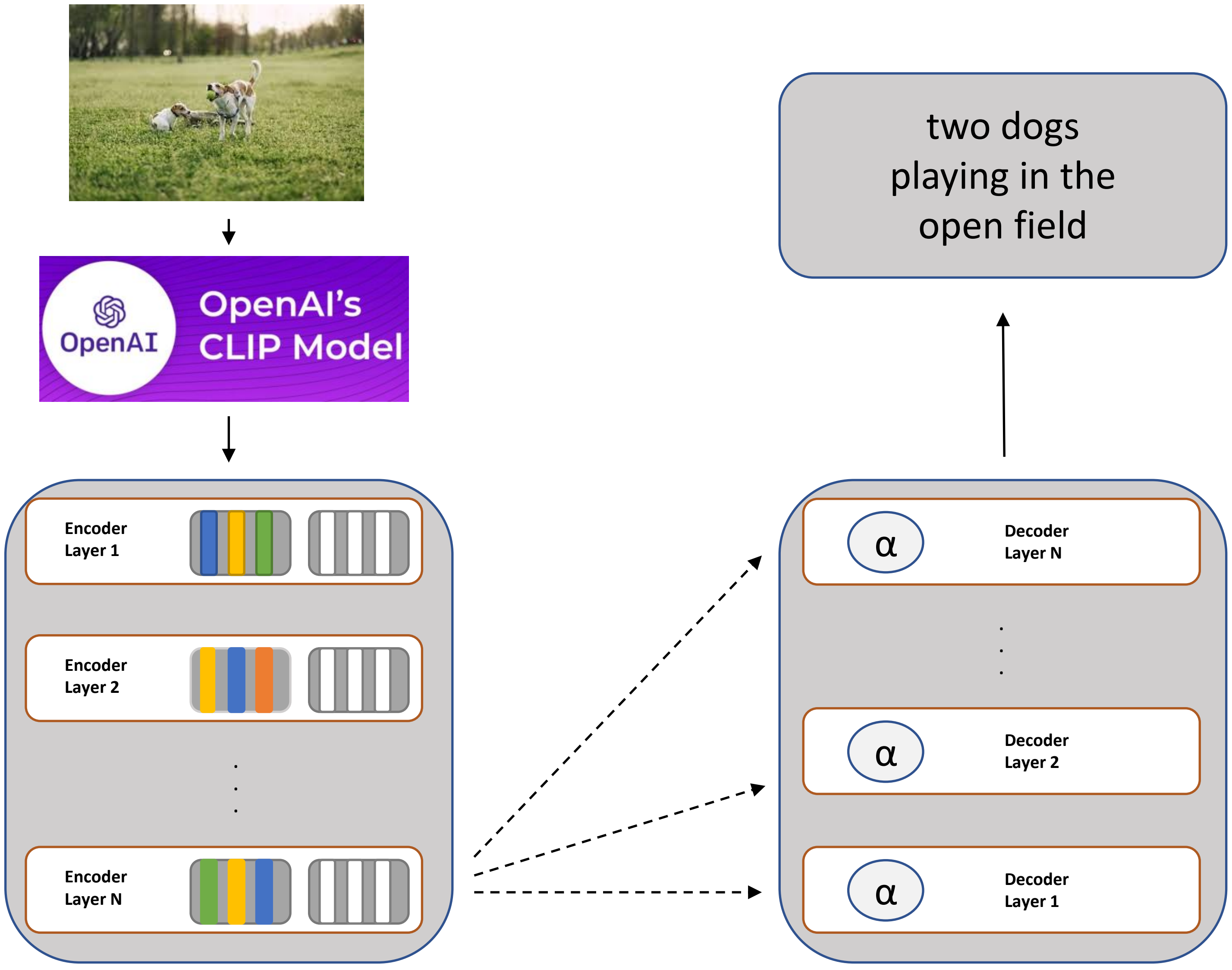
start of the race , athlete , leads

Architecture 1

Main part of the architecture in a Image captioning model



Architecture 2



Memory – Augmented Encoding

Decoding

Training setup of the experiments

	M3 COCO	**M3 finetuned**	**M3 cc_automotive**	**M3 big**
Steps	70k on COCO	50k on COCO 20k on cc_automotive	70k on cc_automotive	70k on COCO
Batch size	25	25	25	50

Evaluation

- During the evaluation we used beam search (5)

	COCO-validation	**COCO-test**	**COCO-automotive**	**CC_automotive-validation**
Number of captions associated with each image	5	1	1	1
Quality of the captions	good	good	good	bad

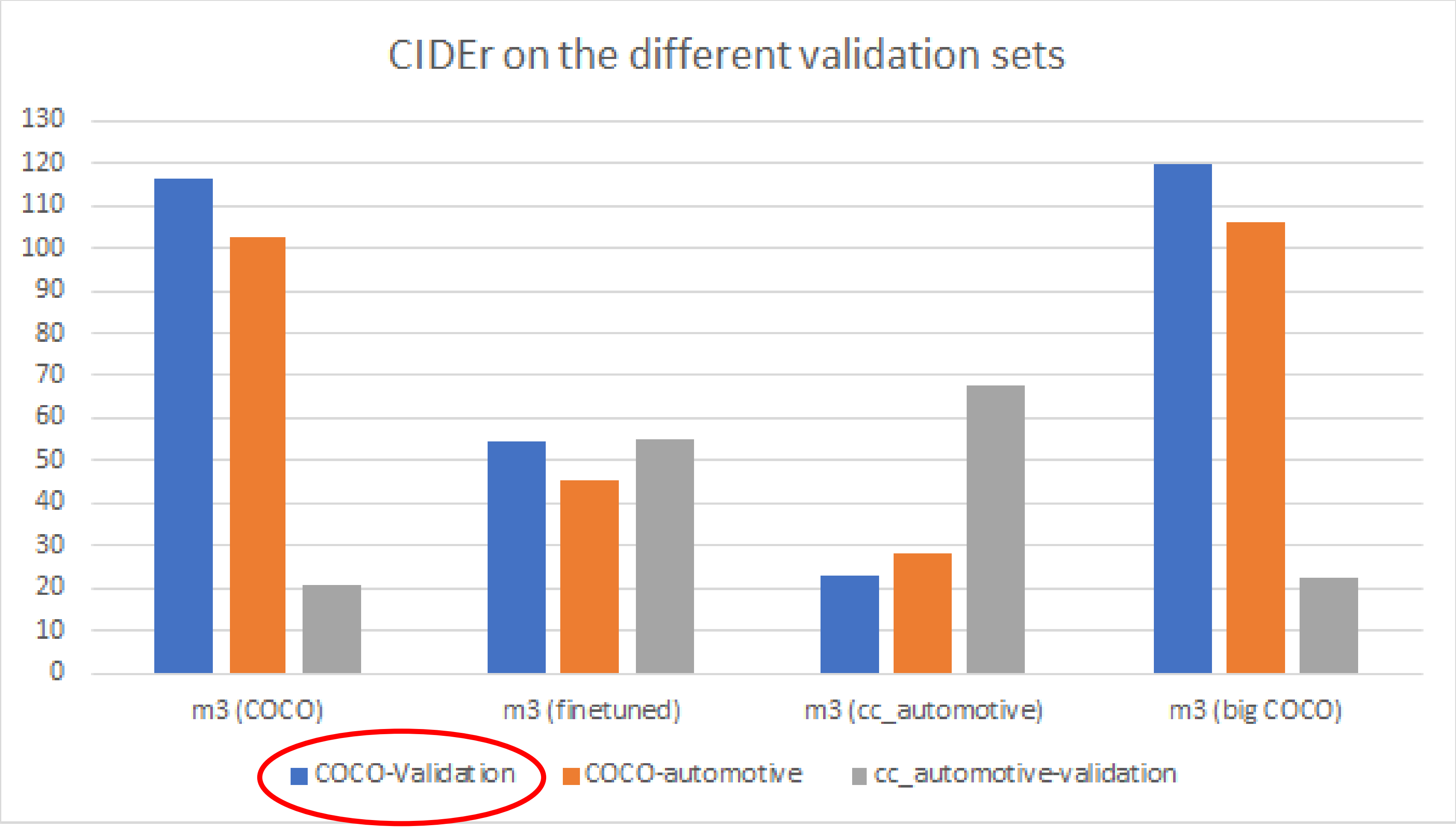
Example taken from COCO

a car sitting at a stop sign in a city.
a vintage sports car at a traffic intersection.
a car is stopped in front of a stop sign
a classic car waiting at a 3-way stop sign.
a car sitting next to a red stop sign in the street.

Example taken from Conceptual Captions

visitors look at cars during the public

Results


CIDEr on the different validation sets
130
120
110
100
90
80
70
60
50
40
30
20
10
0
m3 (COCO)
m3 (finetuned)
m3 (cc_automotive)
m3 (big COCO)
COCO-Validation
COCO-automotive
cc_automotive-validation

3

Results on COCO-validation [1]

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr
m2	81,6	66,4	51,8	39,7	29,4	59,2	129,3
M3_COCO	77,6	62,1	48,3	37,3	27,6	57,4	116,4
M3_finetuned	40,1	29,3	19,9	12,9	16,9	39,4	54,7
M3_cc_automotive	29,2	17,1	9,7	5,1	10,2	27,1	22,8
M3 big	77,6	62,0	48,3	37,4	28,2	57,5	119,7

Results on COCO-validation

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr
m2	81,6	66,4	51,8	39,7	29,4	59,2	129,3
M3_COCO	77,6	62,1	48,3	37,3	27,6	57,4	116,4
M3_finetuned	40,1	29,3	19,9	12,9	16,9	39,4	54,7
M3_cc_automotive	29,2	17,1	9,7	5,1	10,2	27,1	22,8
M3 big	77,6	62,0	48,3	37,4	28,2	57,5	119,7

comparison with benchmark in literature

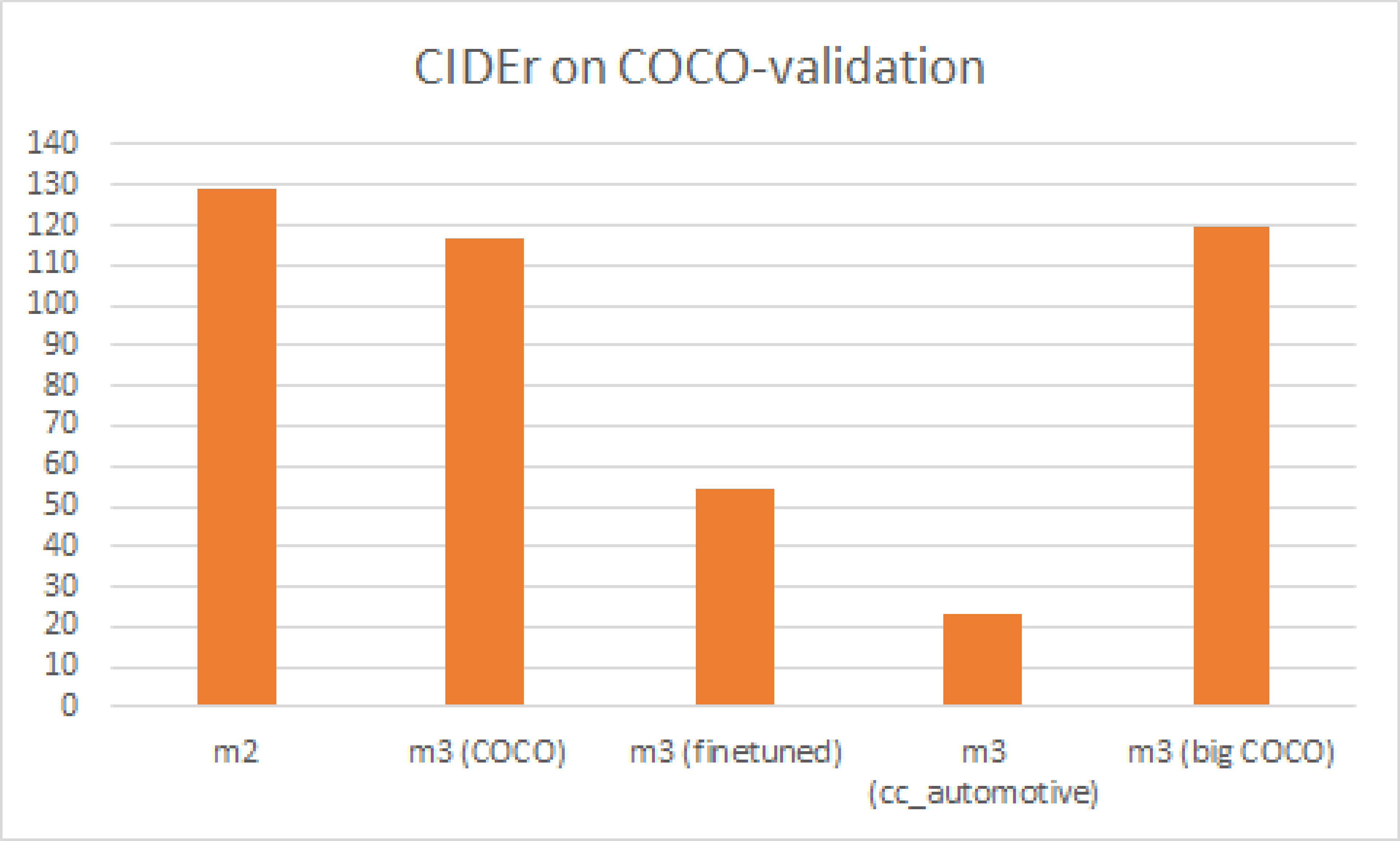
Results on COCO-validation

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr
m2	81,6	66,4	51,8	39,7	29,4	59,2	129,3
M3_COCO	77,6	62,1	48,3	37,3	27,6	57,4	116,4
M3_finetuned	40,1	29,3	19,9	12,9	16,9	39,4	54,7
M3_cc_automotive	29,2	17,1	9,7	5,1	10,2	27,1	22,8
M3 big	77,6	62,0	48,3	37,4	28,2	57,5	119,7

decrease performance
if trained on cc data

Off domain

Results on COCO-validation



CIDEr results 2

	M3-COCO	M3-finetuned	M3-cc_automotive	M3-big
COCO-validation	116,4	54,7	22,8	119,7
COCO-automotive	102,8	45,2	28,2	106,2
CC_automotive	20,7	55,2	67,6	22,6

CIDEr results

	M3-COCO	M3-finetuned	M3-cc_automotive	M3-big
COCO-validation	116,4	54,7	22,8	119,7
COCO-automotive	102,8	45,2	28,2	106,2
CC_automotive	20,7	55,2	67,6	22,6

If we consider the Conceptual Captions eval split, two models outperforms

CIDEr results

	M3-COCO	M3-finetuned	M3-cc_automotive	M3-big
COCO-validation	116,4	54,7	22,8	119,7
COCO-automotive	102,8	45,2	28,2	106,2
CC_automotive	20,7	55,2	67,6	22,6

Out of Domain decreases the performance

CIDEr results

	M3-COCO	M3-finetuned	M3-cc_automotive	M3-big
COCO-validation	116,4	54,7	22,8	119,7
COCO-automotive	102,8	45,2	28,2	106,2
CC_automotive	20,7	55,2	67,6	22,6

On COCO splits perform better on-domain (automotive), considering same ‘type’ of data

Quality of metrics on M3_COCO 3

The table shown the different results if we change the structure of the ground truths, respect to all the metrics

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr
COCO-validation	77,6	62,1	48,3	37,3	27,6	57,4	116,4
COCO-test	38,6	24,7	16,4	11,3	17,6	38,4	115,2

decrease of different metrics, considering **different number of ground truths** → 1 vs. 5

- same quality
- same domain
- same model

Qualitative Results 1

Ground truth:
government agency has received calls a day since boxing day .
Model trained on COCO:
a yellow truck parked in front of a building .
Model trained on cc_automotive:
emergency services at the scene

Ground truth:
visitors look at cars during the public
Model trained on COCO:
a red car parked in front of a crowd of people .
Model trained on cc_automotive:
automotive industry business at show

Qualitative Results 2

Ground truth:
automobile models are the latest vehicles to receive the treatment
Model trained on COCO:
a black car parked in front of a building .
Model trained on cc_automotive:
automobile model on the street

Ground truth:
an image of a man waving from a car .
Model trained on COCO:
a picture of a person in a car .
Model trained on cc_automotive:
cartoon illustration of a man driving a car .

DEMO

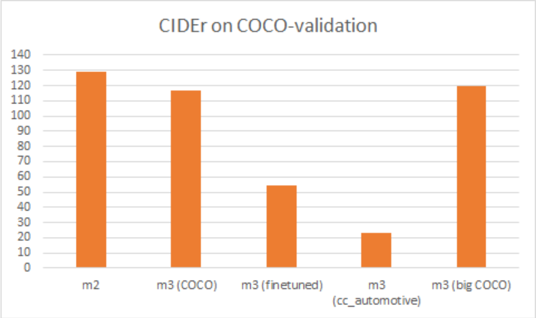
Conclusion

different quality of ground truth	→	training with bad ground truths provides bad output quality
number of captions	→	doesn’t change CIDEr score too much (not the case for the other metrics)
domain of validation data	→	less important with respect to the previous two points

Future Work

To improve the performance we can:

- Increase the dimension of the model
  (as the experiments suggest)



- After the standard training, use Reinforcement Learning

- Meshed and learnable connections [m2]

- Increase the number of training steps

Thank you for your attention!

Paula Klinke

300997@studenti.unimore.it

Federico Cocchi

289824@studenti.unimore.it