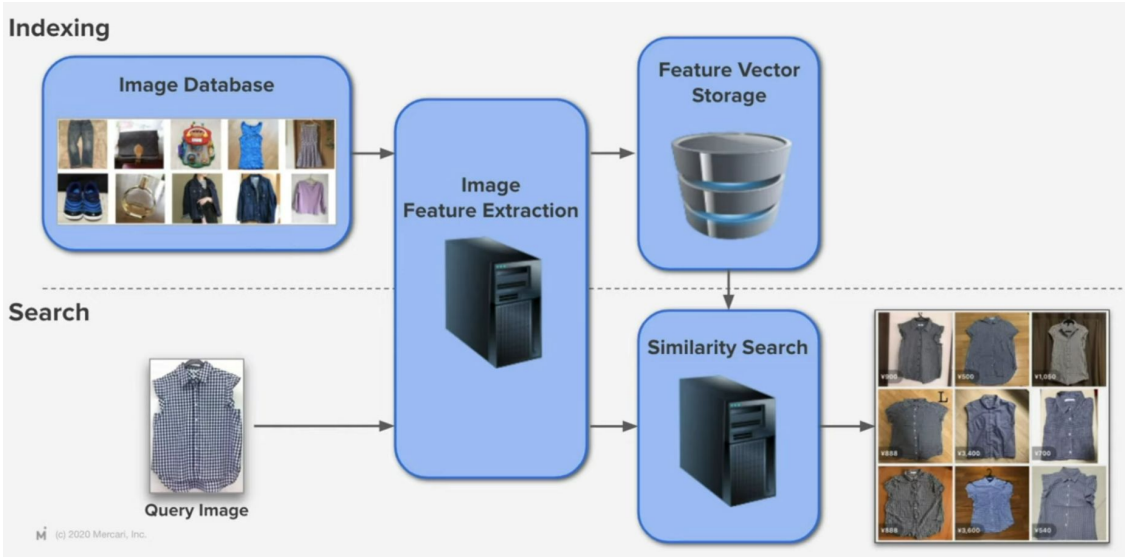
Cross-modal Retrieval

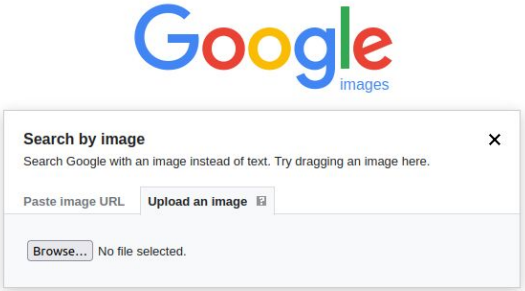
Group members: Jose Manuel López, Alex Martín, Marcos V. Conde

INDEX

- Past weeks recap
- Statement of the problem
- Cross-modal retrieval
- Conclusions week 5
- General conclusions
- Organization of the team

Introduction





CVPR 2020 Image Retrieval Tutorial by Takuma Yamaguchi (Mercari, Inc.)

RECAP: WEEK 1

In the first week, we learned how to implement a model in PyTorch by recreating our best model from M3 module (previously implemented on Keras).

CNN Model from M3.

	Loss	Accuracy
Keras	0.387	0.875
Pytorch	0.782	0.856

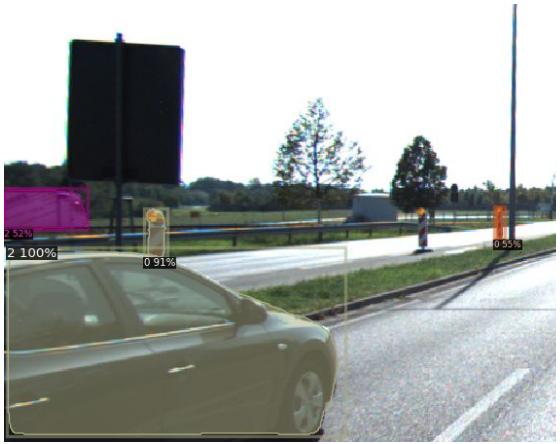
Results from Week 1.

We found our PyTorch implementation(due to errors in the implementation) was about 5 times slower than the Keras one. Results were also better on Keras.

RECAP: WEEK 2

The second week consisted on running detectron2 and evaluate Faster R-CNN and Mask R-CNN over KITTI-MOTS dataset. We evaluated models with COCO weights for segmentation and detection tasks.

- Mask R-CNN had better results than Faster R-CNN
- ROI batch size haves impact in the mAP.
- Small objects were the most difficult to detect or segment



Faster R-CNN:
R_50_FPN_1x

ROI batch size	mAP
256	57.160
512	58.038
1024	55.470

Mask R-CNN:
R_50_FPN_1x

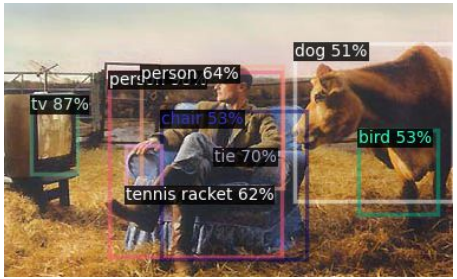
ROI batch size	mAP detection	mAP segmentation
64	58.384	45.572
124	57.239	46.363
256	56.064	44.988

RECAP: WEEK 3

The third week consisted on running fine-tuned models over artificial/hard scenarios. We run inference over out of context images, transplanting objects in images and isolate object from context.

Task a)

The out of context images forced the model to have unexpected behaviors, but overall could perform better than expected.

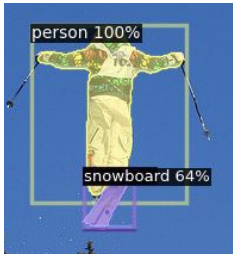


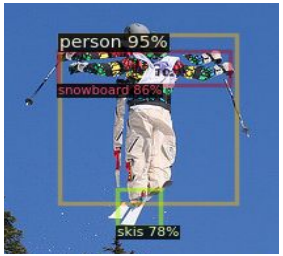
Task b)

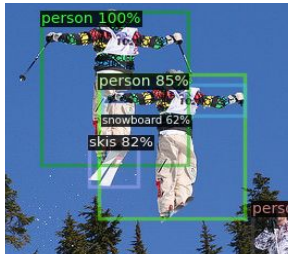
When transplanting objects to another image, we mostly saw that the models had trouble to detect them if there was some similar texture or colors in the destination image or could wrongly classify the object. If the object is not similar to anything in the image, then it can correctly detect most of the time.

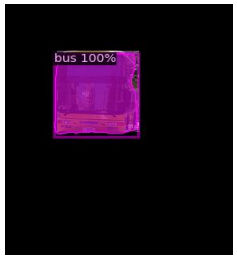
RECAP: WEEK 3

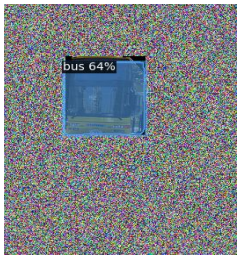
The third week consisted on running fine-tuned models over artificial/hard scenarios. We run inference over out of context images, transplanting objects in images and isolate object from context.

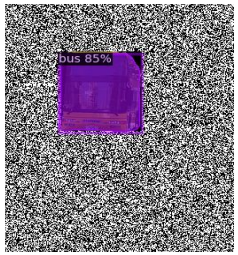












Task c)

When transplanting an object over the image, we found that in many cases it actually helped the model to correctly detect other objects which in the original image weren't detected, although in general it was less accurate.

Task d)

When the object is isolated from the rest of the image, the model doesn't seem to have strange behaviors, but when adding noise at the background we can observe a worse performance of the models.

RECAP: WEEK 4

The fourth week consisted on running image retrieval with different embeddings methods: we tested ResNet, Siamese and Triplet Networks.

Later, we used KNN and FAIS to perform the retrieval with the different embeddings.

Results on the different approaches tested

Changing the backbone for the embedding had a very negative impact in the performance. The Triplet network shows a great improvement with respect to the Siamese and having much fewer parameters than in the pretrained model the MAP isn't much lower.

For the Siamese and Triplet we used embedding of 10 feature vectors which caused the downgrade in the MAP.

Retrieval methods

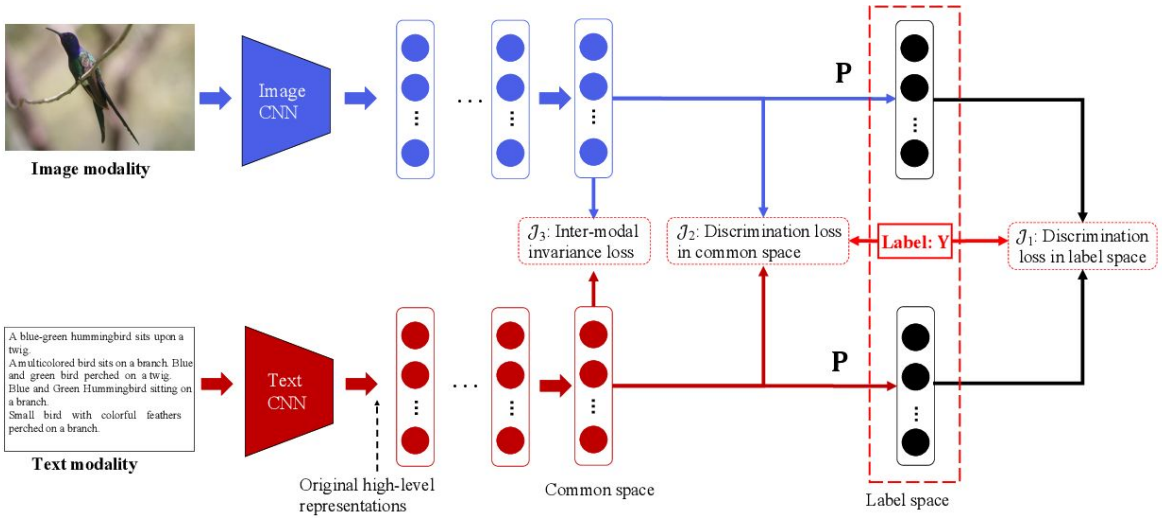
The methods compared, KNN and FAIS, with all the embedding that were used show that FAIS is slightly better.

Test	ResNet	Siamese	Triplet
MAP	0.78	0.18	0.6

Results from Week 4.

STATEMENT OF THE PROBLEM

Cross-modal retrieval aims to enable flexible retrieval across different modalities. The core of cross-modal retrieval is how to measure the content similarity between different types of data.



Example of a CNN based Cross-Modal Retrieval with images and text. Source

WEEK 5: CROSS-MODAL RETRIEVAL

DATASET: Flickr 30k

Dataset with 31014 images and 5 captions per image where images and captions focus on people involved in everyday activities and events. Split of training, validation and test used from Karpathy et al.

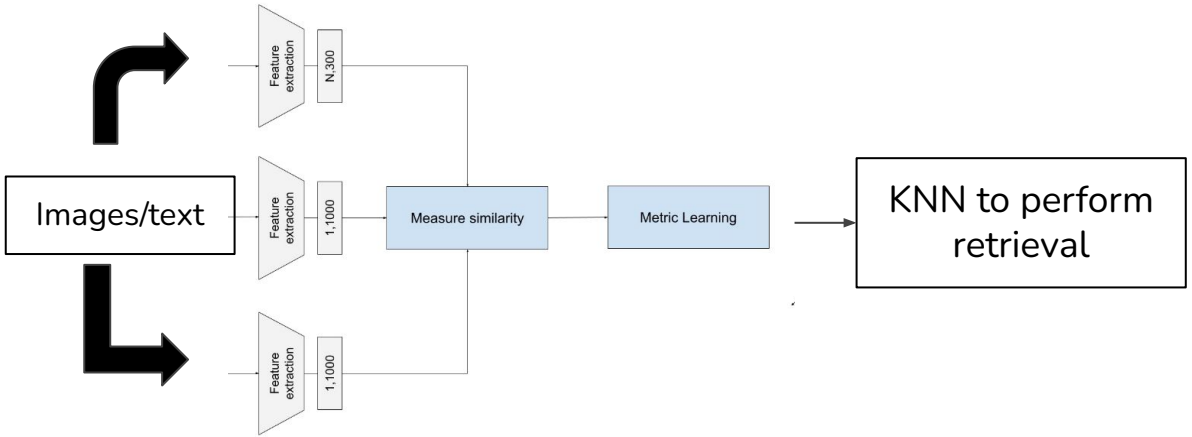
Captions
A mother decides to take her child on a piggyback ride outside their apartment complex.
A baby boy in a blue and white striped shirt is sitting on his mother's shoulders.
A mother and her young son enjoying a beautiful day outside.
A young woman is giving a baby a ride on her shoulders.
A woman gives a small child a piggyback ride.

WEEK 5: CROSS-MODAL RETRIEVAL

Metric learning

We used 2 different embeddings for images (VGG and FasterRCNN) and 2 for text (FastText and BERT.

To perform the retrieval, we used the same pipeline as in the previous week.



WEEK 5: CROSS-MODAL RETRIEVAL

Metric learning

Triplet with a neural network for the image embeddings and another one for the text embeddings. Small networks (3 and 2 layers) which use a combination of Linear and Relu.

Loss: Triplet loss
Optimizer: Adam
Learning Rate: 0.001
Other information: used learning rate scheduler and 50 epochs of training used.

Main focus of our experiments:

- Research the effect of the variation in the output size and the text aggregation method to the accuracy

WEEK 5: CROSS-MODAL RETRIEVAL

Image to text retrieval

Text aggregation method: Sum

Embedding size	Train Loss	Validation Loss	Accuracy
256	0.186	0.179	0.085
512	0.201	0.211	0.096
1024	0.229	0.237	0.096
2048	0.264	0.271	0.095
4096	0.267	0.278	0.092

Text aggregation method: Mean

Embedding size	Train Loss	Validation Loss	Accuracy
256	0.146	0.163	0.103
512	0.145	0.172	0.099
1024	0.138	0.169	0.087
2048	0.137	0.158	0.102
4096	0.152	0.206	0.093

WEEK 5: CROSS-MODAL RETRIEVAL

Image to text retrieval
Qualitative Results

Predictions

A young female student performing a downward kick to break a board held by her Karate instructor.
The man is dressed in all black and has a tattoo on his left arm.
A woman talks to her newborn child.
A lady pushes a stroller, behind a child, riding a tricycle, while another child walks alongside.
A woman at a fish market is looking at the fish that are frozen.

WEEK 5: CROSS-MODAL RETRIEVAL

Image to text retrieval
Qualitative Results

Predictions

A girl kicking a stick that a man is holding in taekwondo class.

Five people wearing winter clothing, helmets, and ski goggles stand outside in the snow.

A mountaineer about to descend down a mountain with a blue helmet on.

A person in a black coat pulling a kid out of the road.

A young woman is laying in the sun with her face covered by a purple scarf.

WEEK 5: CROSS-MODAL RETRIEVAL

Image to text retrieval
Qualitative Results

Girl about to kick a piece of wood in half while karate instructor holds it
A young girl swimming in a pool
Little white dog wearing a leash jumping after a red ball.
A woman holds a young boy who has a wooden spoon in his mouth.
A Heavy machine lifting up a worker.

WEEK 5: CROSS-MODAL RETRIEVAL

Text to image retrieval with VGG

Text aggregation method: Sum

Embedding size	Train Loss	Validation Loss	Accuracy
256	0.275	0.309	0.344
512	0.319	0.367	0.358
1024	0.489	0.567	0.344
2048	0.558	0.771	0.368
4096	0.832	1.704	0.334

Text aggregation method: Mean

Embedding size	Train Loss	Validation Loss	Accuracy
256	0.352	0.366	0.234
512	0.380	0.553	0.231
1024	0.507	0.831	0.238
2048	0.645	1.269	0.230
4096	0.849	2.050	0.220

WEEK 5: CROSS-MODAL RETRIEVAL

Text to image retrieval with VGG
Qualitative Results

Query text:
A small child kissing a cat on the kitchen counter

Ground truth

Prediction

Query text: A girl in martial arts class is kicking a dummy

Ground truth

Prediction

WEEK 5: CROSS-MODAL RETRIEVAL

Text to image retrieval with VGG
Qualitative Results

Query text:
The toddler in blue is looking away from his red ball

Ground truth

Prediction

Query text: An offensive player running with a football while a football player tries to stop him during a football game

Ground truth

Prediction

WEEK 5: CROSS-MODAL RETRIEVAL

Text to image retrieval with Faster RCNN

Text aggregation method: Sum

Embedding size	Train Loss	Validation Loss	Accuracy
256	0.162	0.210	0.422
512	0.160	0.184	0.427
1024	0.167	0.170	0.420
2048	0.186	0.252	0.419
4096	0.222	0.300	0.412

Text aggregation method: Mean

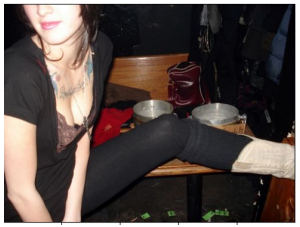
Embedding size	Train Loss	Validation Loss	Accuracy
256	0.210	0.229	0.344
512	0.199	0.212	0.342
1024	0.206	0.266	0.346
2048	0.242	0.313	0.317
4096	0.275	0.396	0.324

WEEK 5: CROSS-MODAL RETRIEVAL

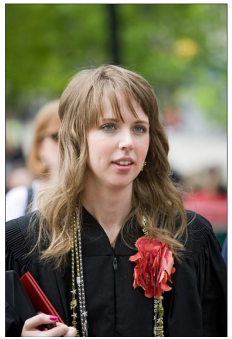
Text to image retrieval with Faster RCNN
Qualitative Results

Query text:
A girl dressed in black is posing for the camera

Ground truth

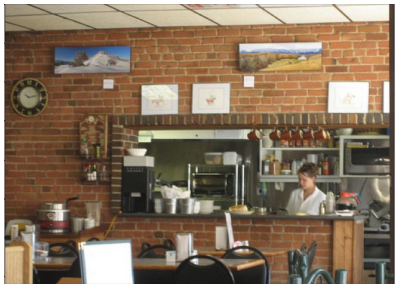


Prediction



Query text: A young woman in white working in a professional kitchen

Ground truth



Prediction

WEEK 5: CROSS-MODAL RETRIEVAL

Text to image retrieval with Faster RCNN
Qualitative Results

Query text:
A child is rollerblading with a lot of pads

Ground truth

Prediction

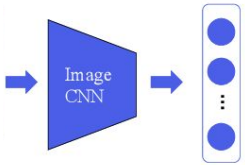
Query text: A brown dog is panting hard on grass during a sunny day

Ground truth

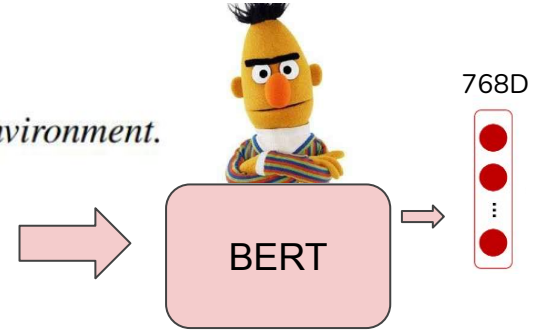
Prediction

WEEK 5: CROSS-MODAL RETRIEVAL

Text to image retrieval with BERT



Gray haired man in black suit and yellow tie working in a financial environment.
A graying man in a suit is perplexed at a business meeting.
A businessman in a yellow tie gives a frustrated look.
A man in a yellow tie is rubbing the back of his neck.
A man with a yellow tie looks concerned.



WEEK 5: CROSS-MODAL RETRIEVAL

Text to image retrieval with BERT

Gray haired man in black suit and yellow tie working in a financial environment.
A graying man in a suit is perplexed at a business meeting.
A businessman in a yellow tie gives a frustrated look.
A man in a yellow tie is rubbing the back of his neck.
A man with a yellow tie looks concerned.

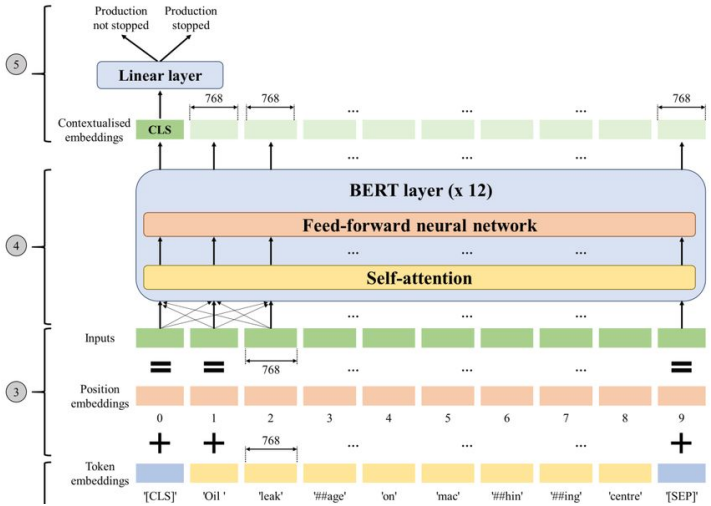
[101, 1037, 2158, 2007, 1037, 3756, 5495, 3504, 4986, 102] =tokens
[0, 0, 0, 0, 0, 0, 0, 0, 0, 0] = token types
[1, 1, 1, 1, 1, 1, 1, 1, 1, 1] = mask

[CLS] a man with a yellow tie looks concerned [SEP]

WEEK 5: CROSS-MODAL RETRIEVAL

Text to image retrieval with BERT

- We use *“bert-base-uncased”* model and tokenizer from HuggingFace.
- Tokenizer “max_length = 64”, we also add padding and truncate.
- For each caption (5) we obtain the tokens and masks from the tokenizer
- We obtain a 768-dimensional embedding from BERT’s CLS token (from the “last_hidden_states”)
- Problem? using 32bits precision, the size for 31k samples is +400mb, and it takes several hours to calculate.



WEEK 5: CROSS-MODAL RETRIEVAL
Conclusions

Image to text vs text to image retrieval

We expected to have a similar performance in both tasks, but we had a better performance in the text-to-image than in the image-to-text

Qualitative results

Analyzing them, we can see that many errors that the retrieval model does are understandable and ‘close’ to what the model should retrieve.

Text aggregation method

From the two tested methods we saw that the adding the embedding of the words in a sentence had better results in text-to-image while performing the mean resulted in better results in the image-to-text. The best method will be specific to each case

M5: General conclusions

Pytorch
Although in the beginning, it was more challenging than Keras in terms of more advanced Deep learning methods, PyTorch is better.

Object detection and segmentation
Main problem of the models tested are the small objects. The dataset used in the training haves a greater impact in the performance than in the classification task.

Retrieval
Metric learning is an efficient method, based on a solid approach and gives also the opportunity to perform cross-modal retrieval, although this may not have the best performance

Datasets
The many datasets that we've been dealing with in this module have shown us the impact that it can have in the results of a model. We've seen it with the out-of-context task as well as this week that the model was 'understandably' failing in many cases.

Organization

Main tasks

Alex: started the implementation of the code, editing of slides and summaries
Josep: debugging and continue the implementation, writing of slides
Marcos: debugging and continue the implementation, writing of report and slides

	Coding	Slides	Report
Alex	70%	20%	10%
Josep	55%	25%	20%
Marcos	50%	20%	30%