Master in
Computer Vision
Barcelona

M5 Project: Cross-modal Retrieval

Week 5

Cross-modal Retrieval

Rubèn Pérez Tito
rperez@cvc.uab.cat

Ernest Valveny
ernest@cvc.uab.cat

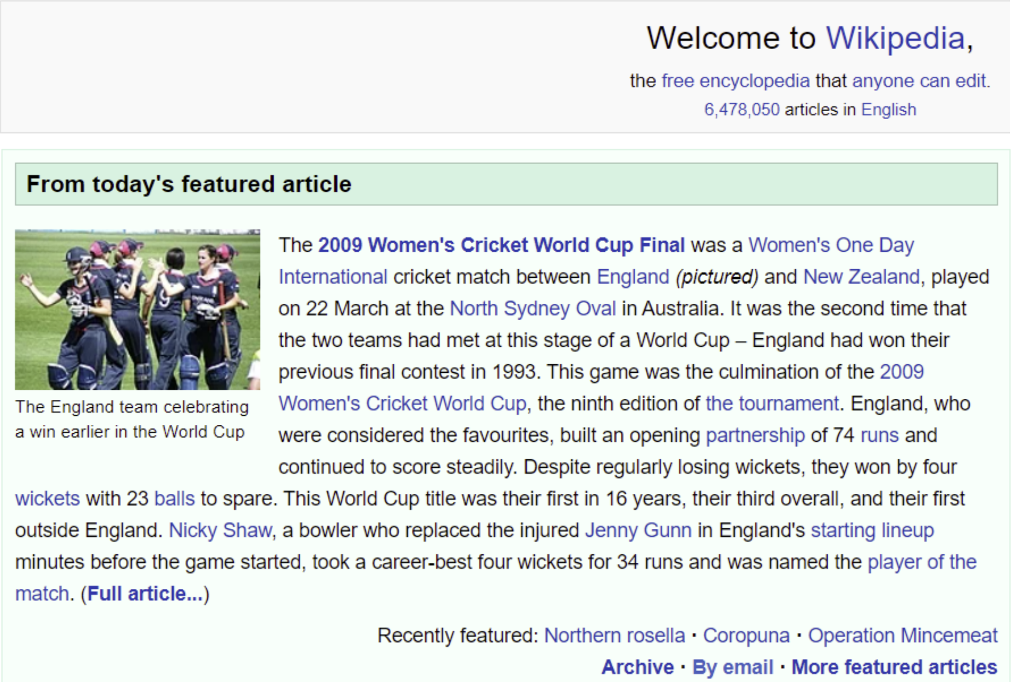
M5 Project Stages and Schedule

Week 1 March 6-12	**P1: Introduction to Pytorch - Image Classification**
Week 2 March 13-19	**P2 & P3: Object Detection, Recognition and Segmentation**
Week 3 Marh 20 - 26	
Week 4 March 27 – April 2	**P4: Image Retrieval**
EASTER	
Week 5 April 17 - 23	**P5: Cross-modal Retrieval** **Deliverable: Report on object Detection and Segmentation, first version**
Week 6 April 24	**Deliverable: Presentation** **Deliverable: Report on object Detection and Segmentation, final version**

M5 – Natural Language

Humans communicate through some form of language either by text or speech which conveys **high semantic information**. To make interactions between computers and humans, computers need to understand natural languages used by humans

- Used in many ways:
  - Communicate information (article).
  - Describe an image (caption).
- It requires a specific processing.



Welcome to Wikipedia,

the free encyclopedia that anyone can edit.
6,478,050 articles in English

From today's featured article

The England team celebrating a win earlier in the World Cup

The **2009 Women's Cricket World Cup Final** was a Women's One Day International cricket match between England *(pictured)* and New Zealand, played on 22 March at the North Sydney Oval in Australia. It was the second time that the two teams had met at this stage of a World Cup – England had won their previous final contest in 1993. This game was the culmination of the 2009 Women's Cricket World Cup, the ninth edition of the tournament. England, who were considered the favourites, built an opening partnership of 74 runs and continued to score steadily. Despite regularly losing wickets, they won by four wickets with 23 balls to spare. This World Cup title was their first in 16 years, their third overall, and their first outside England. Nicky Shaw, a bowler who replaced the injured Jenny Gunn in England's starting lineup minutes before the game started, took a career-best four wickets for 34 runs and was named the player of the match. (**Full article...**)

Recently featured: Northern rosella · Coropuna · Operation Mincemeat
Archive · **By email** · **More featured articles**

M5 – Natural Language Processing (NLP)

Natural Language Processing (NLP) is a branch of artificial intelligence that analyzes, models and generates language that humans naturally use, in order to interact with them both in written and spoken contexts.

NLP Tasks:

- Machine translation
- Text Summarization
- Text classification
- Sentiment Analysis
- Dialog systems (chatbots → ChatGPT)
- …

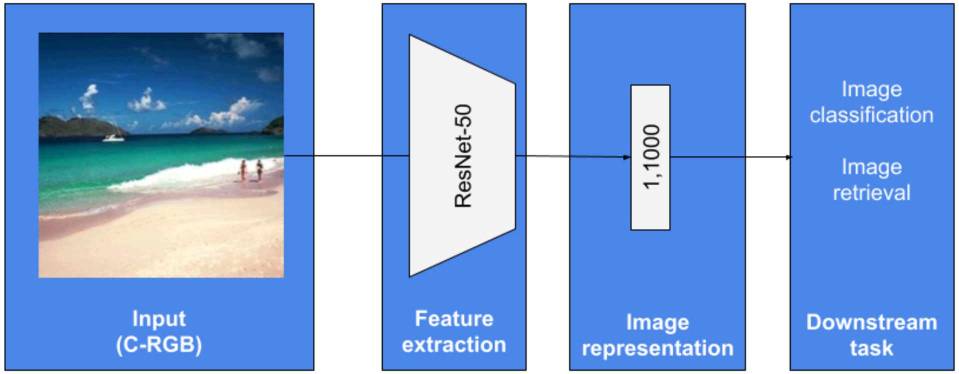
Involved in CV tasks (multimodal):

- Image captioning
- Visual question answering (VQA)
- **Cross-modal retrieval**
  - **Image-to-text**
  - **Text-to-image**
- Visual Dialog (GPT4)

In this project we see a very tiny part of this

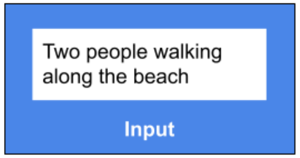
M5 – Working with Natural Language

Common Computer Vision pipeline:



$W \cdot H \cdot C \in \mathbb{R}$

Natural language input?



M5 – Word embeddings

We need to find a way to represent a string in a way that neural networks can process.

- Not learned:
  - One-hot vectors from a fixed vocabulary.
  - Pyramidal Histogram of Characters (PHOC)
  - …
- Learned:
  - Word2Vec
  - Global Vectors (GloVe)
  - FastText
  - BERT
  - …

Word embeddings

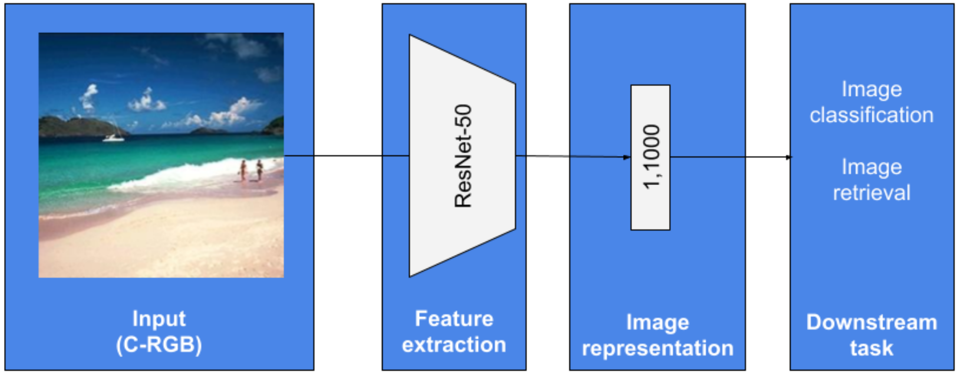
Each embedding has its own properties and therefore its pros and cons.

All learned embeddings start by mapping the words represented in one-hot vectors to the learned representation.

- When a word is not included in the one-hot vector dictionary, it's called Out of vocabulary (OOV) word.
  - Minimizing the impact of OOV words is one of the main challenges in NLP.

M5 – Working with Natural Language

Image stream pipeline:



Input
(C-RGB)

ResNet-50

Feature extraction

1,1000

Image representation

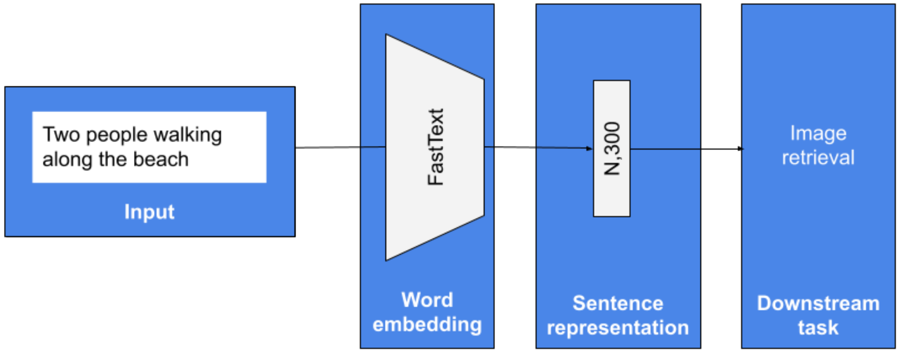
Image classification

Image retrieval

Downstream task

$W \cdot H \cdot C \in \mathbb{R}$

Language steam pipeline:



Two people walking along the beach

Input

FastText

Word embedding

N,300

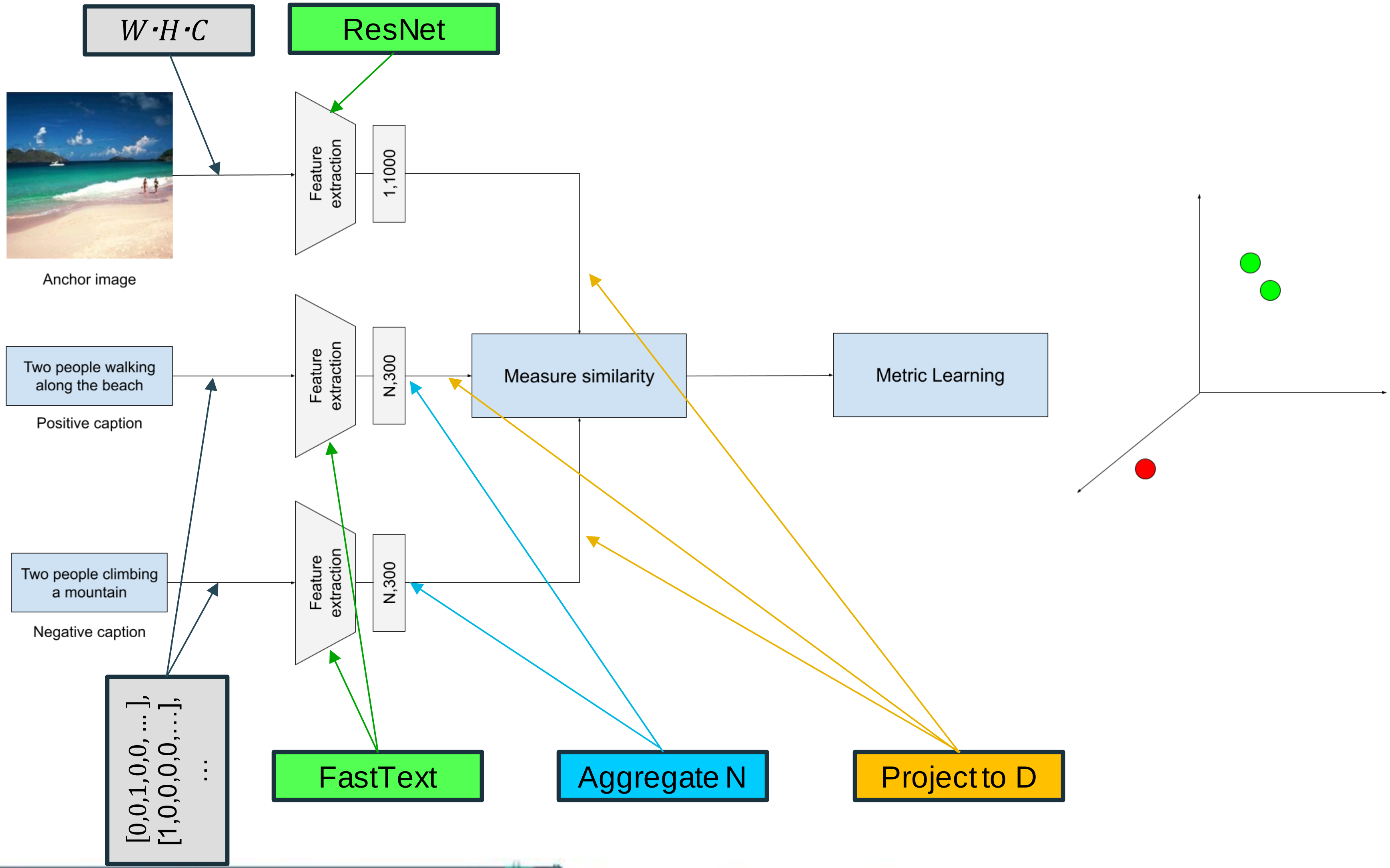
Sentence representation

Image retrieval

Downstream task

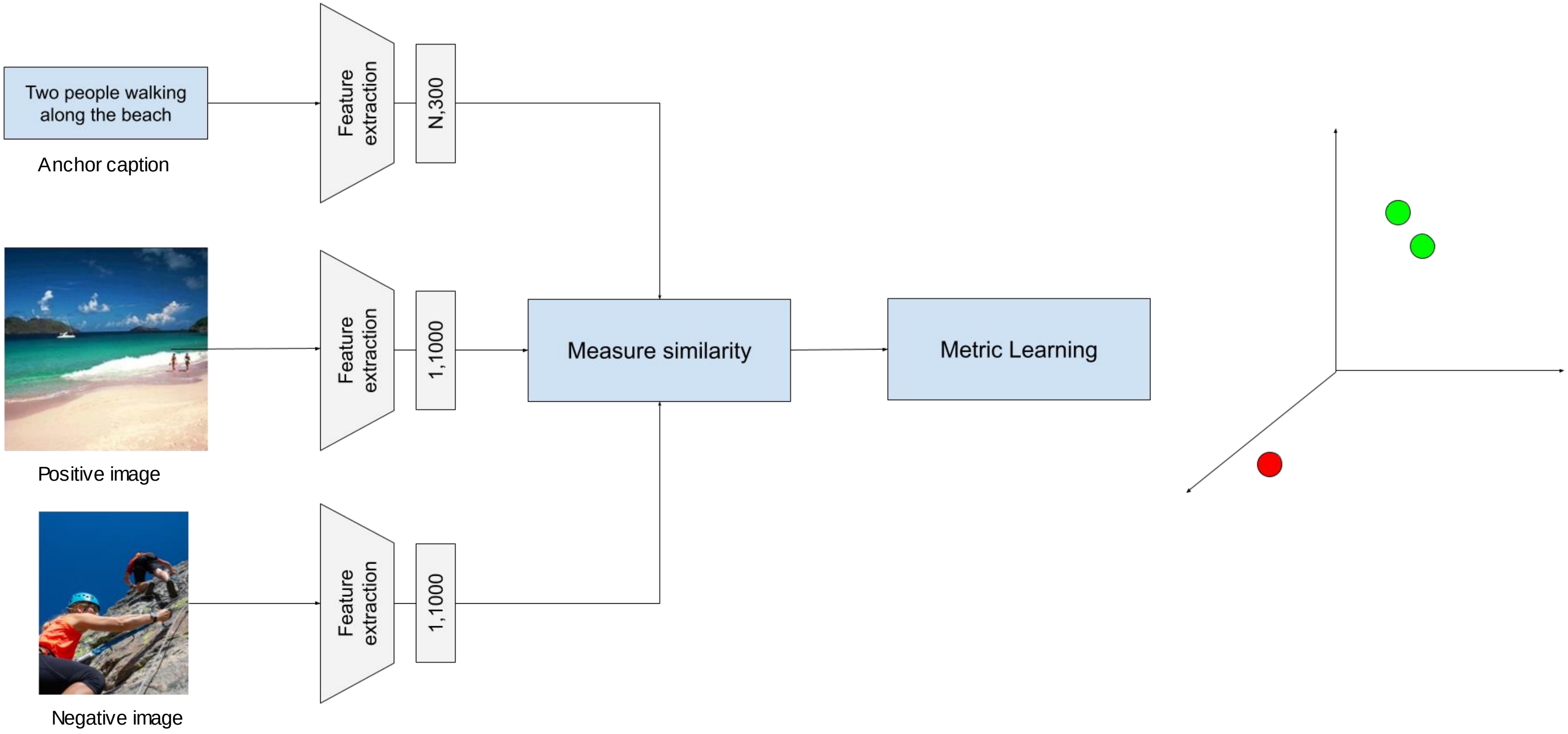
M5 – Cross-modal retrieval

Image-to-text retrieval: The objective is to retrieve a correct caption given an image.



M5 – Cross-modal retrieval

Text-to-image retrieval: The objective is to retrieve a correct image given a caption.



M5 – Literate Models for computer vision

AIDA Course: Literate Models for Computer Vision (link)

- Detection and Recognition approaches and comparison of current SotA OCR systems
- Language representation (**embeddings**)
- Fine-grained Image Classification
- **Cross-modal retrieval**
- Scene Text Visual Question Answering
- Document Visual Question Answering
- Demo session (fine-grained image classification)

M5 – P5 Tasks

Week 5: Cross-modal Retrieval

Tasks

a. Implement basic Image-to-text retrieval.
b. Implement basic Text-to-image retrieval.
c. Use BERT embedding as Text feature extractor.
d. Review the report with the provided feed-back.
e. Prepare final presentation

Deliverable (for next week)

- **Github** repository with readme.md (code explanation & instructions)
- Presentation with all items listed in the tasks.
- **Final** version of the **Report** on overlaf.

M5 – P5 Tasks

Dataset: COCO 2014

- /home/mcv/datasets/COCO/
  - captions_train2014.json
  - captions_val2014.json

- train_captions[’annotations’]

```
[
    {’image_id’: 318556, ’id’: 48, ’caption’: ’A very clean and well…’},
    {’image_id’: 116100, ’id’: 67, ’caption’: ’A panoramic view of…’},
    …
]
```

M5 – P5 Tasks

FastText

- Pip install fasttext
- /home/mcv/m5/fasttext_wiki.en.bin
- model = fasttext.load_model(`"model_filename.bin"`)

- `Word in model` 🡐 To know if the Word is OOV or not.
- `Model[word]` 🡐 To get the representation
  - **Lowercase!**

M5 – P5 Tasks

Task (a): **Implement basic Image-to-text retrieval.**

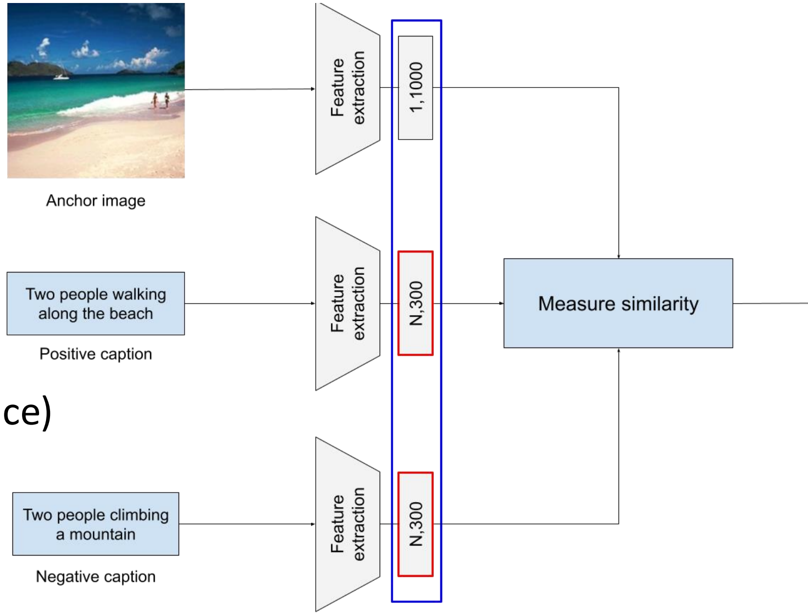
- Anchor Image.
- Positive caption: Caption corresponding to the anchor image.
- Negative caption: Any other caption.

- Image stream (choose one):
  - ResNet / Faster R-CNN / Mask R-CNN

- Language stream:
  - FastText

- Choose measure similarity procedure (Euclidean distance)
  - Project features to the same space (blue).

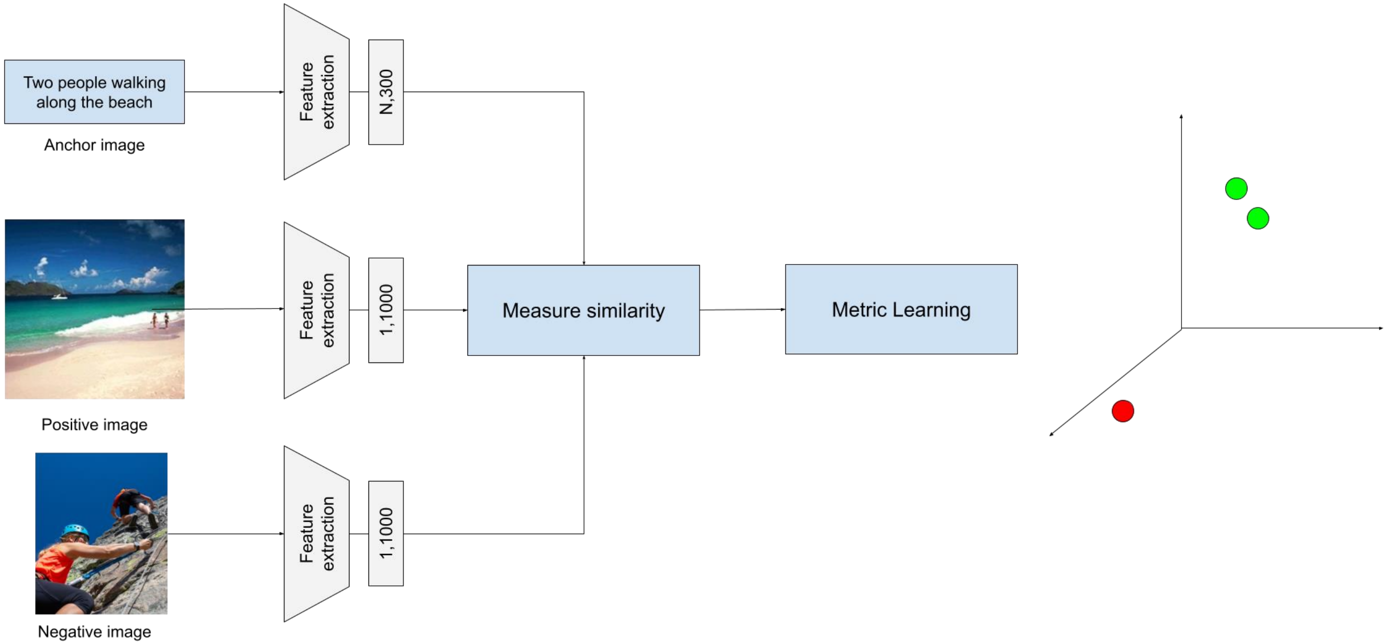
- Choose textual aggregation scheme (red).



M5 – P5 Tasks

Task (b): **Implement basic Text-to-image retrieval.**

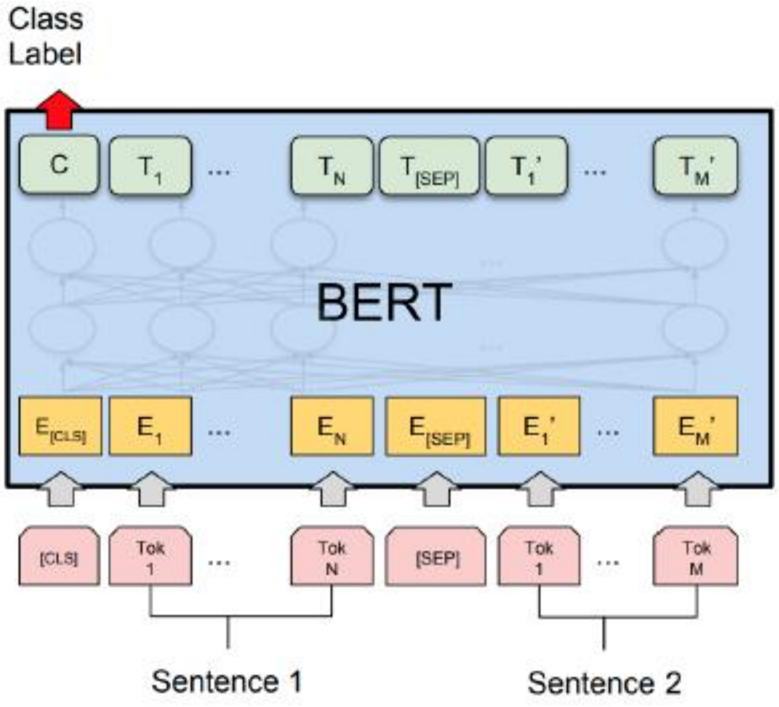
- Anchor caption.
- Positive image: Image corresponding to the anchor caption.
- Negative image: Any other image from the training set.



M5 – P5 Tasks

Task (c): **Use BERT embedding as Text feature extractor.**

- Huggingface library BERT model



M5 – P5 Tasks

For all tasks a, b and c:

- If you face memory problems especially during retrieval.
  - Use a subset of the evaluation as the database.
    - **Detail the final setting you use**.

M5 – P5 Tasks

Task (d): **Review the report with the provided feed-back.**

- What if we have a 10 in all the report deliveries?

Task (e): **Prepare the final presentation.**

- Oral presentation of up to 10 minutes
- Include one slide with internal organization of the group and coordination of the tasks.
- Describe in detail one of the projects P2, P3, P4 or P5
  - **Different format!**
- Include a summary of the work done in the rest of projects (one slide per week)
- Include a slide with conclusions defining valuable lessons/interesting findings during module 5.
- All group member must participate in the oral presentation.

M5 – P5: Image Retrieval

Due date

24th of April, Monday, before 10:00 AM