Master in Computer Vision Barcelona

M5 Project: Cross-modal Retrieval

Week 5

Cross-modal Retrieval

Rubèn Pérez Tito
rperez@cvc.uab.es

Ernest Valveny
ernest@cvc.uab.es

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.
- Involved in CV tasks:
  - Image captioning
  - Visual question answering (VQA)
  - **Cross-modal retrieval**
    - **Image-to-text**
    - **Text-to-image**

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, understands and generates language that humans naturally use, in order to interact with them both in written and spoken contexts.

- In this project we see a very tiny part of this.

Other NLP tasks:

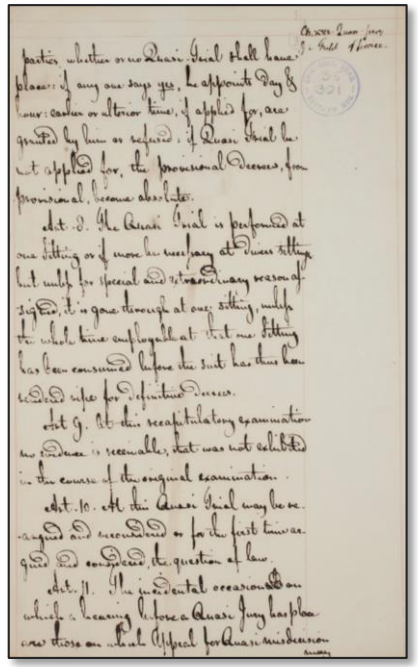
- Machine translation
- Text Summarization
- Text categorization
- Sentiment Analysis
- Dialog systems (chatbots)
- …

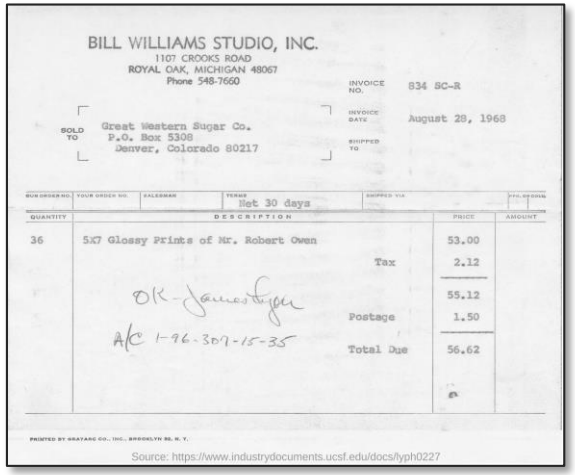
M5 – Natural Language

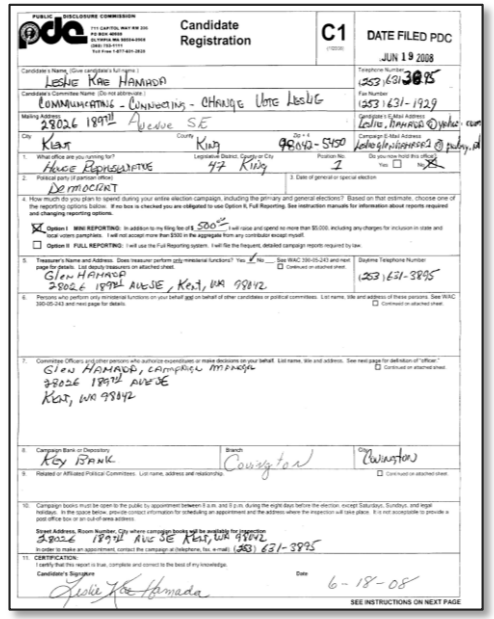
2 main sources of language:

- Text present in images.
  - Scenes or scanned documents.
  - Optical Character Recognition (OCR).

- Free text
  - Paragraphs (articles), dialogs, questions, answers, captions.
  - Text is given.

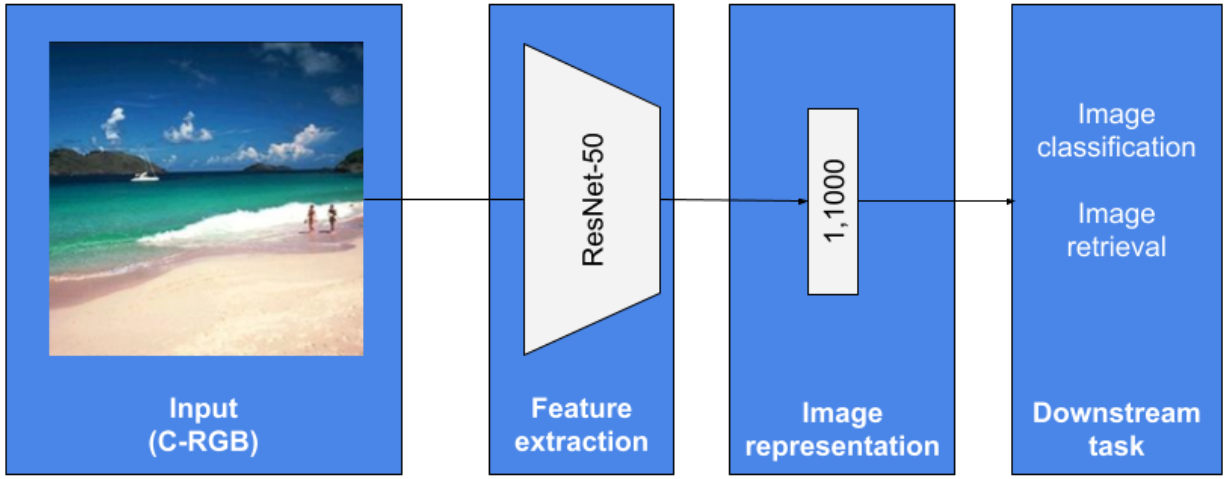






M5 – Word embeddings

Common Computer Vision pipeline:



Natural language input?

M5 – Word embeddings

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

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

Word embeddings

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

M5 – Literate Models for computer vision

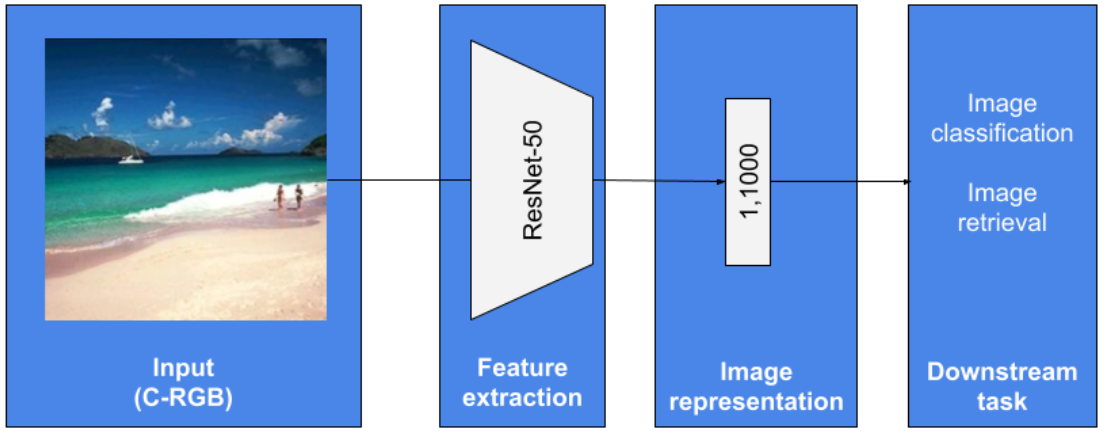
AIDA Course: Literate Models for Computer Vision (link)

- Monday April 4
  - Detection and Recognition approaches and comparison of current SotA **OCR** systems
  - Language representation (**embeddings**)
  - Fine-grained Image Classification

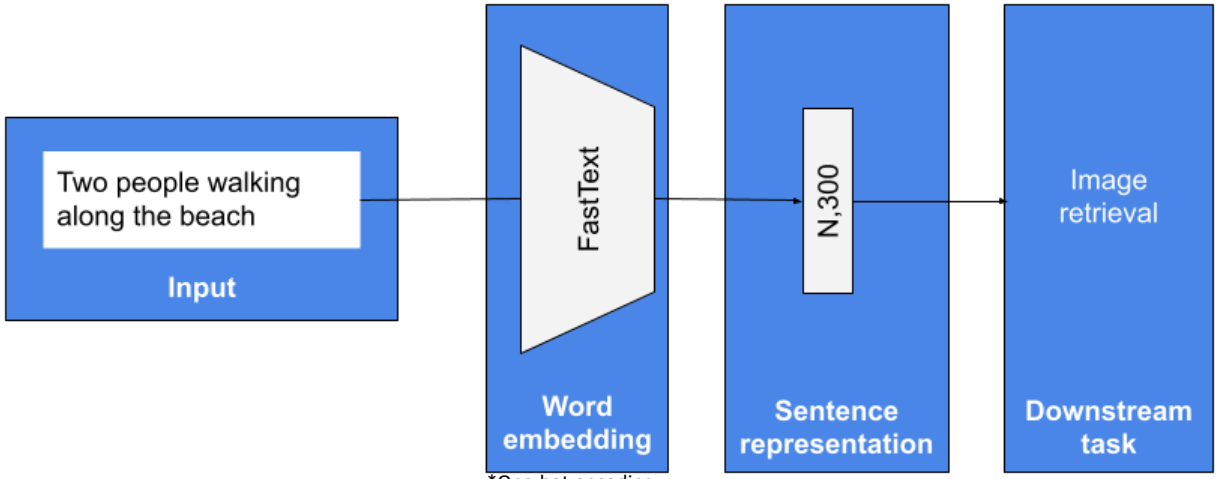
- Wednesday April 6
  - **Cross-modal retrieval**
  - Scene text Visual Question Answering
  - Document Visual Question Answering
  - Demo session (fine-grained image classification)

M5 – Word embeddings

Image stream pipeline:

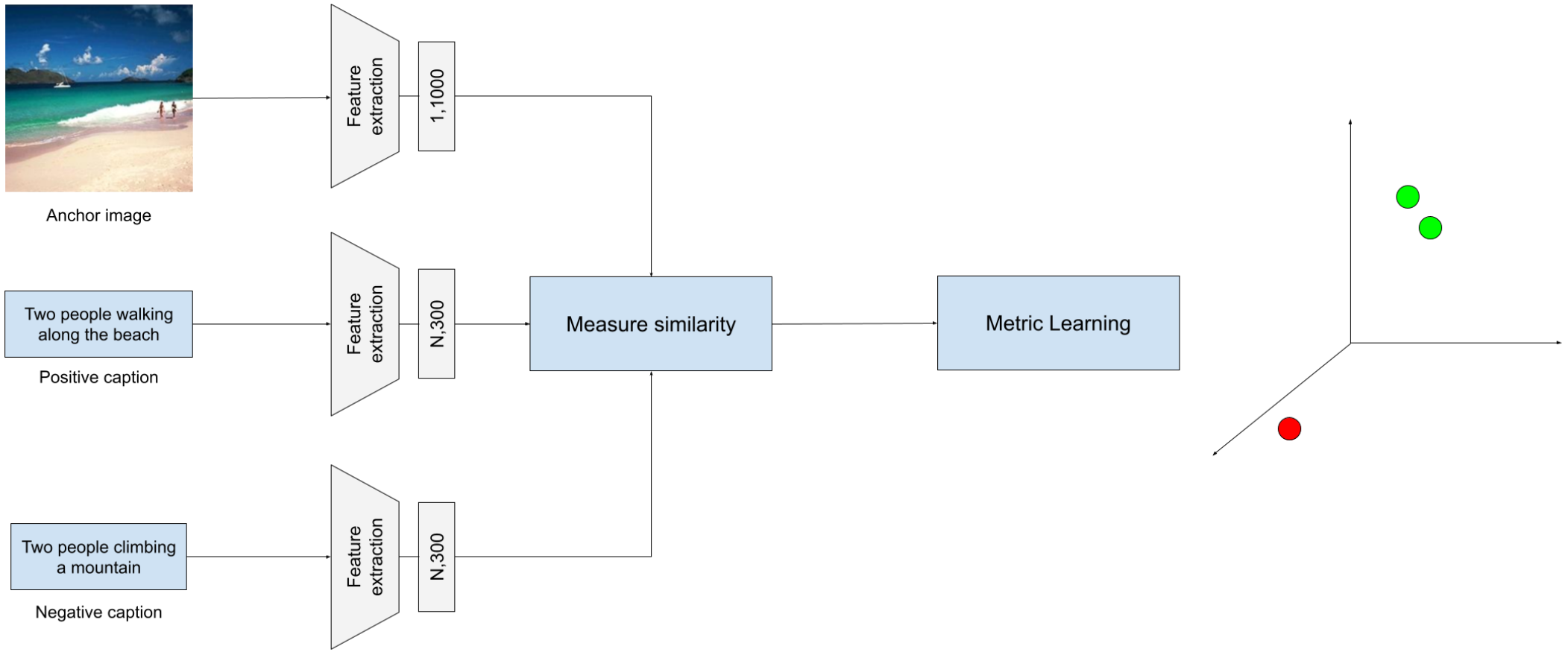


Language stream pipeline:



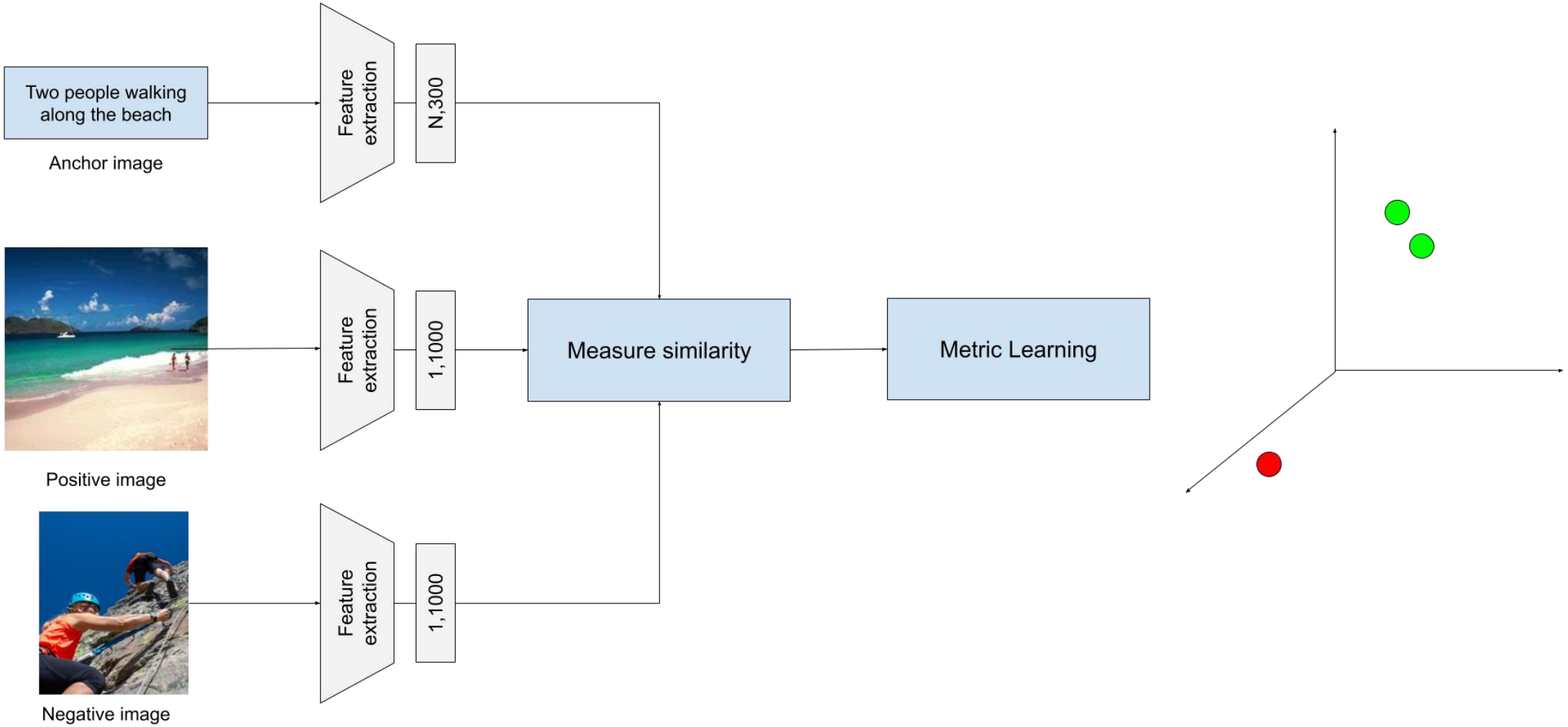
M5 – Cross-modal retrieval

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 – Cross-modal retrieval

Main research interest is how we **combine features** from different sources.

Fusion scheme:

- Feature aggregation:
  - Concatenation, sum, average.
  - Attention
    - Classic attention (dot product)
    - Self-attention (transformers)
- Where the aggregation takes place:
  - Early fusion
  - Late fusion

M5 – Cross-modal retrieval

Main research interest is how we **combine features** from different sources.

Pre-computed features.

- Given a fixed inputs representations, how the interaction between those can be improved?
- Pros:
  - Reduce the memory usage and speeds-up the training procedure.
- Cons:
  - The goal of the final task might help to better understand/model your inputs (boost of performance).
- Middle point:
  - Load the N last layers of the feature representation model and fine-tune it.

Week 5. Cross-modal retrieval

Details on tasks, deliverables, and marks for this week

Week 1	**Introduction to Pytorch - Image Classification**
Week 2	**Object Detection, Recognition and Segmentation I**
Week 3	**Object Detection, Recognition and Segmentation II**
	Object Classification, Detection and Segmentation Report
Week 4	**Image Retrieval**
Week 5	**Cross-modal Retrieval**
Week 6	**Image and Cross-modal retrieval Report**
	Final Presentation

M5 Project: Goals per week

Week 5: Cross-modal retrieval

Goals

(a) Implement basic Image-to-text retrieval.
(b) Implement basic Text-to-image retrieval.
(c) Use Faster R-CNN as Image feature extractor.
(d) Use BERT embedding as Text feature extractor.
(e) (Optional) LSTM as aggregation for textual features.
(f) Finish writing the retrieval report.
(g) Prepare final presentation

Marks

(C) Achieve (a, b), (f, g) goals
(B) Achieve (a - c), (f, g) goals
(A) Achieve (a - d), (f, g) goals

Deliverable (for next week)

- **Github** repository (code explanation & instructions)
- **Final presentation**
- **Report** on overleaf about image and cross-modal retrieval.

M5 – Cross-modal Retrieval

Dataset

- Flickr 30K
  - 31K images. Official split from Karpathy et al. link
  - 5 captions per image.
    - Database from train (images/captions).

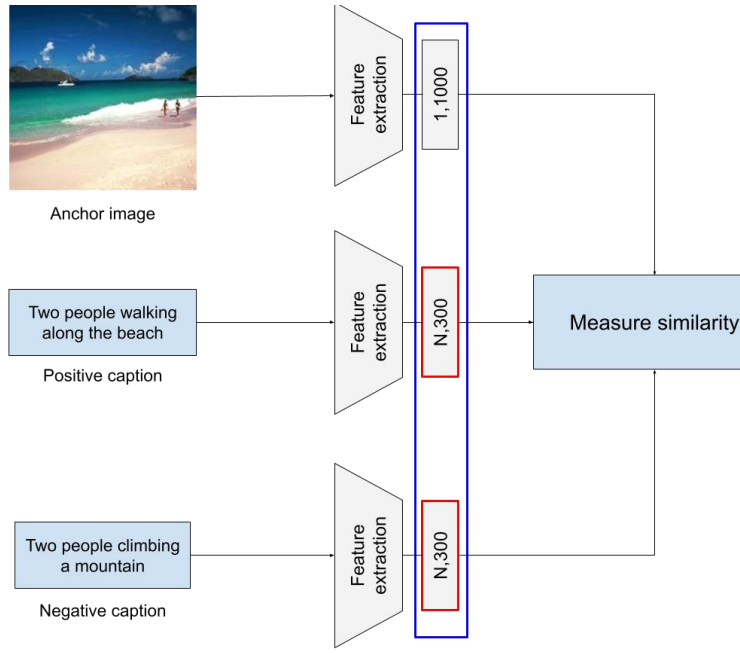
  - /home/mcv/datasets/Flickr30k

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.

M5 – Cross-modal Retrieval

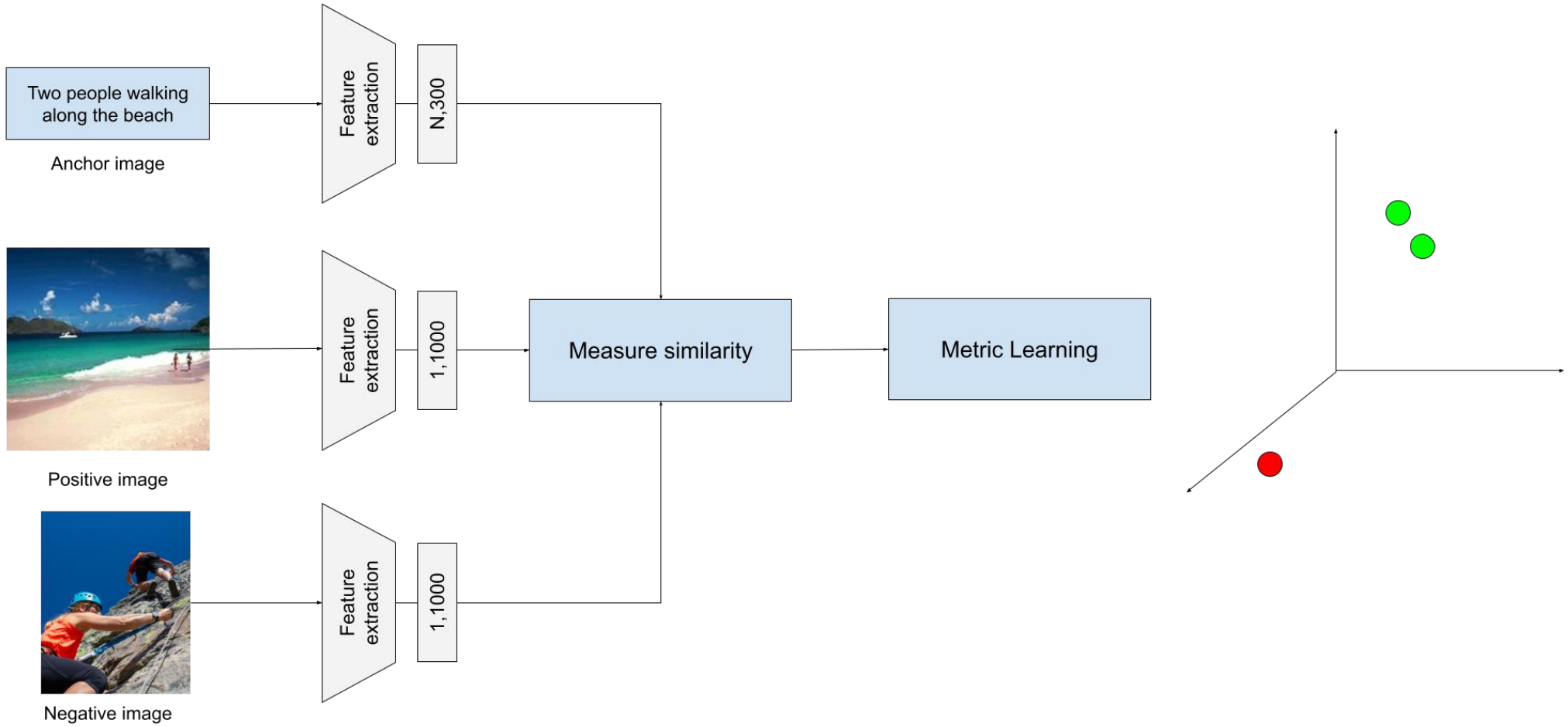
Task (a): **Implement basic Image-to-text retrieval.**

- Image stream:
  - You will find pre-computed image features from VGG-16 in **vgg_feats.mat**.
    - Shape: (4096, 31014)
- Language stream:
  - You will find pre-computed FastText features in **fasttext_feats.npy**.
    - Shape: (31014, 5, W, 300)
- Choose measure similarity procedure (Euclidean distance)
  - Project features to the same space (blue).
- Choose textual aggregation scheme (red)



M5 – Cross-modal Retrieval

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



M5 – Cross-modal Retrieval

Task (c): **Use Faster R-CNN to extract visual features.**

- Choose only one of the retrieval schemes (image-to-text or text-to-image).
- Replace the VGG pretrained features for a Faster R-CNN.
  - Choose image aggregation scheme.

- (Optional): Load the weights of the last Faster R-CNN layer to fine-tune it during training.

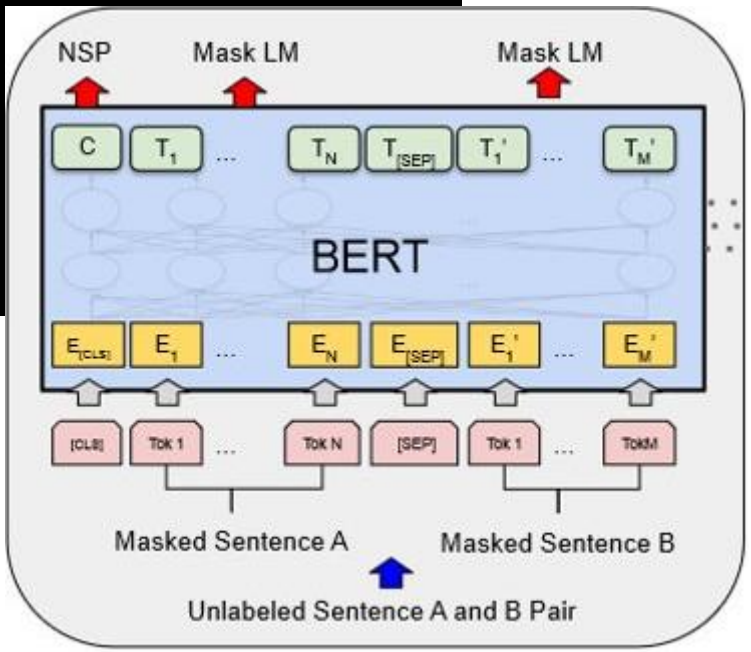
M5 – Cross-modal Retrieval

Task (d): **Use BERT embedding for textual features.**

- Hugginface transformers library.
- Understand BERT tokenizer (CLS token) and model output.

```
!pip install transformers
from transformers import BertTokenizer, BertModel
bert_tokenizer = BertTokenizer.from_pretrained('bert-base-uncased')
bert_model = BertModel.from_pretrained('bert-base-uncased')

sample_text = "Hello, nice to meet you"
inputs = bert_tokenizer(sample_text)
last_hidden = bert_model(**inputs)['last_hidden_state']
```



M5 – Cross-modal Retrieval

Task (e): **(Optional) LSTM as aggregation for textual features for FastText.**

- Pytorch Long short-term memory (LSTM).
  - Use LSTM last hidden state as caption representation.
  - Might need to include <bos> and <eos> tokens.

M5 – Cross-modal Retrieval

Task (f): **Write a final version of the paper**

- Extend related work section with textual embeddings
- Include a section on methodology describing the frameworks for image retrieval and cross-modal retrieval
- Complete the section on experiments including cross-modal retrieval experiments
- Include Abstract, Introduction and Conclusions sections
- Compact the paper into a 6 pages two-column paper using CVPR template
  - You need to include the most relevant findings on the main paper
  - You may include other experiments done as supplementary material

M5 – Cross-modal Retrieval

Task (g): **Prepare the final oral presentation**

- Oral presentation of up to 10 minutes
  - Include one slide with internal organization of the group and coordination of the tasks.
  - Summarize the whole work done during the whole project
  - Main focus on the last task about cross-modal retrieval
  - Summarize main findings from the rest of tasks
  - Include a slide with conclusions defining valuable lessons/interesting findings during module 5
  - All group member must participate in the oral presentation

M5 – Cross-modal Retrieval

- **Code** on Github project
- Prepare the final **presentation**
- **Overleaf** link on your Github

Due date: Monday 25th April before 10:00 AM