# Course 6: Multimodal Foundation Models

# Summary

**Last session**

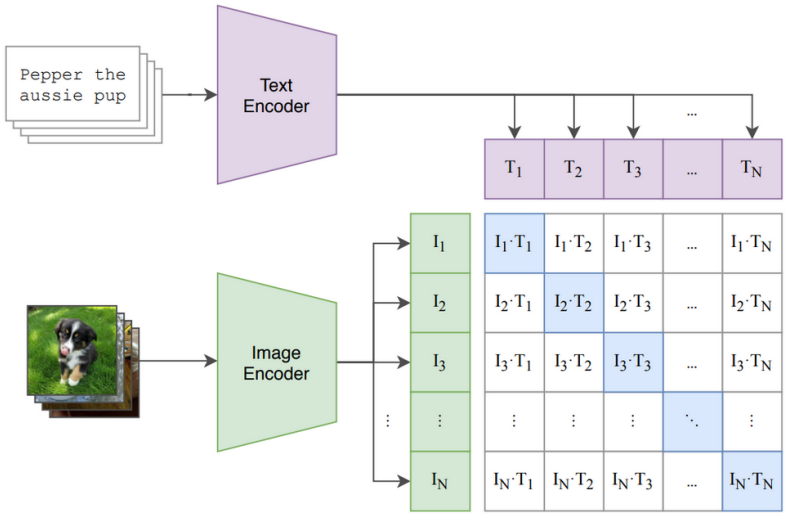
1. What is a Foundation Model
2. Self Supervised Learning
3. Some examples of Foundation Models

**Today's session**

- What is a Multimodal Foundation Model
- Contrastive Learning
- CLIP
- CLAP

# What is a Multimodal Foundation Model?

- Trained on internet-scale multimodal dataset
- Training task si not straightforward (SSL, contrastive pretext tasks)
- Generic feature extractors, Multipurpose
- Generalization is not a problem anymore! All is about **particularization**

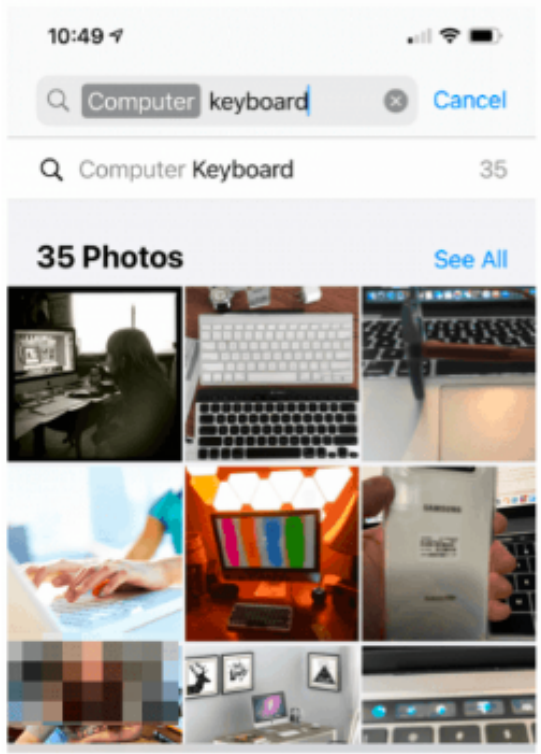



*CLIP:* `https://arxiv.org/abs/2103.00020v1`

# New multimodal tasks

Multimodal Foundation Models enable a wide range of tasks beyond traditional unimodal applications like classification and segmentation.

- Image Retrieval

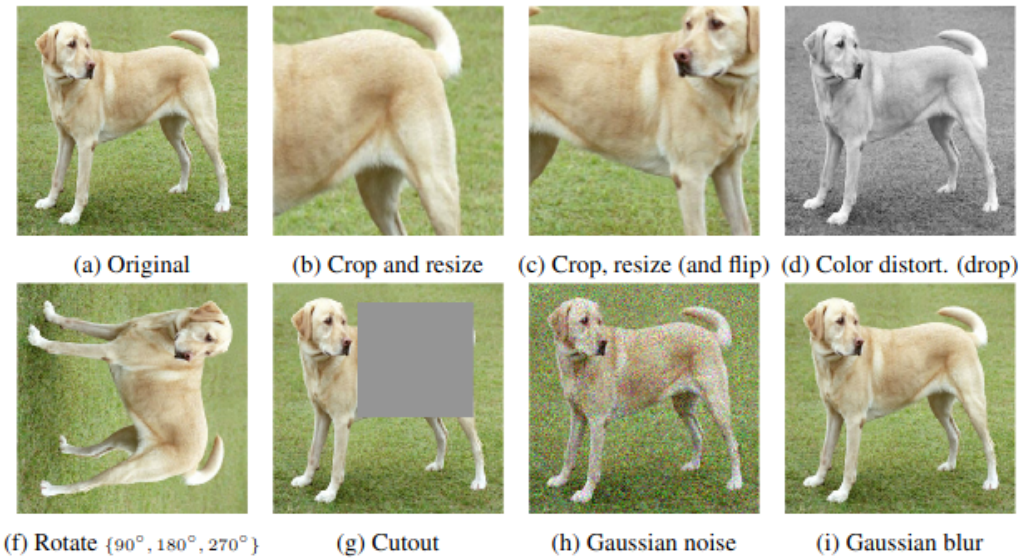


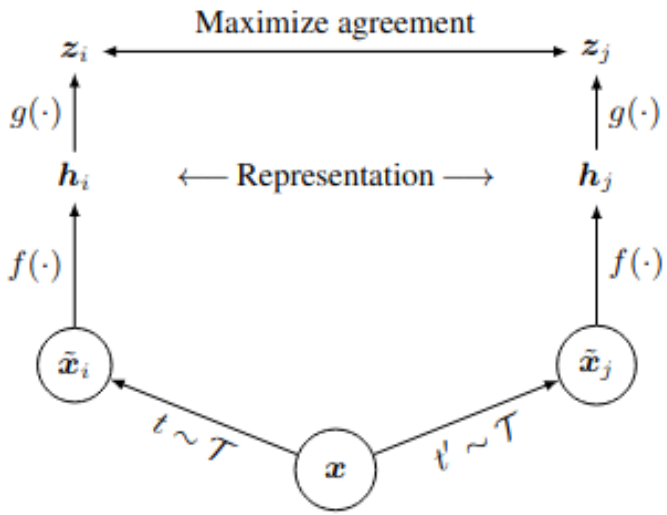

- Text to Image Generation

# Self Supervised Contrastive Learning

## Types of SSL approaches

- **Masked Input Modeling**: Predicting missing part of the input
- **Contrastive Learning**: Pulling together similar representations and pushing apart dissimilar ones




*SimCLR:* `https://arxiv.org/pdf/2002.05709`, *BYOL:* `https://arxiv.org/pdf/2006.07733`

# Self Supervised Contrastive Learning

## Types of SSL approaches

- **Masked Input Modeling**: Predicting missing part of the input
- **Contrastive Learning**: Pulling together representations of the same class in different modalities

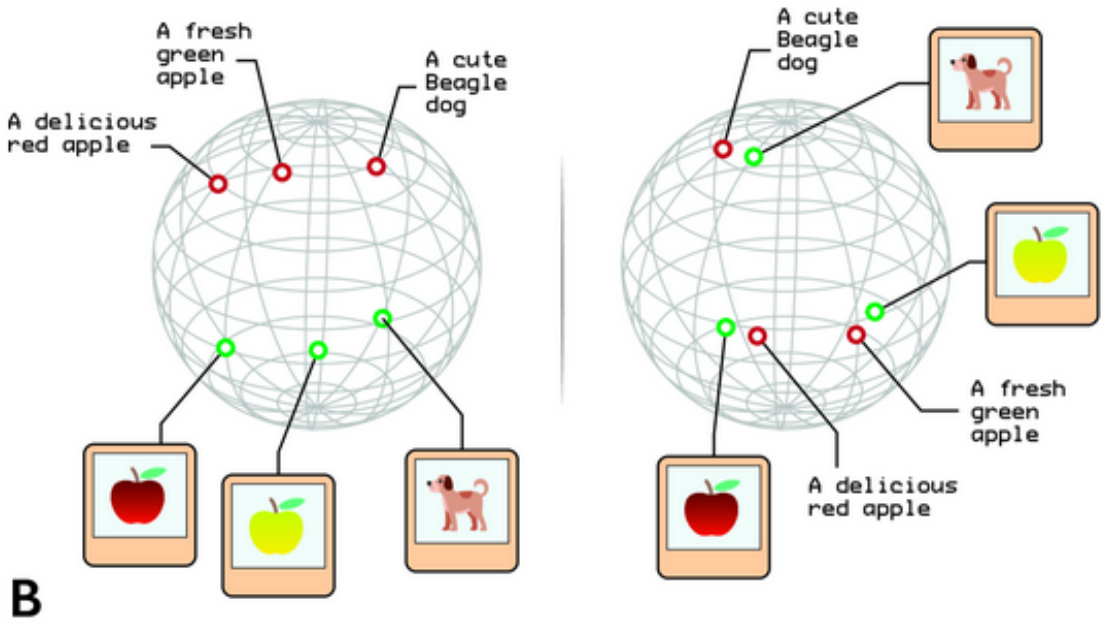



*Image from:* `https://arxiv.org/html/2406.17639v1`

# CLIP: Contrastive Language-Image Pre-Training

## Main Ingredients

- Pretrained on a dataset of internet image-text pairs
- Two separate encoders: one for images (Vision Transformer) and another for text (Large Language Model).
- Contrastive Learning: The model learns by maximizing the similarity between image and text embeddings for matching pairs

The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.

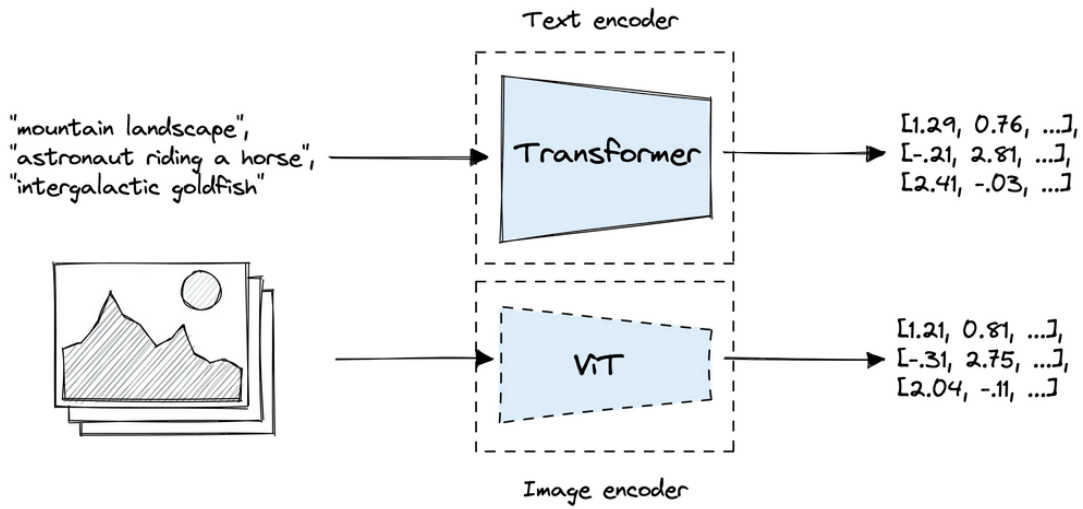
*Zero-Shot CLIP:* `https://www.pinecone.io/learn/series/image-search/zero-shot-image-classification-clip/`,

*Tutorial by Yassir Bendou:*
`https://github.com/brain-bzh/clip/tree/71ff8784d9c37ed279e660a77aede0ffeab69515?tab=readme-ov-file`

# CLIP: Contrastive Language-Image Pre-Training

## Main Ingredients

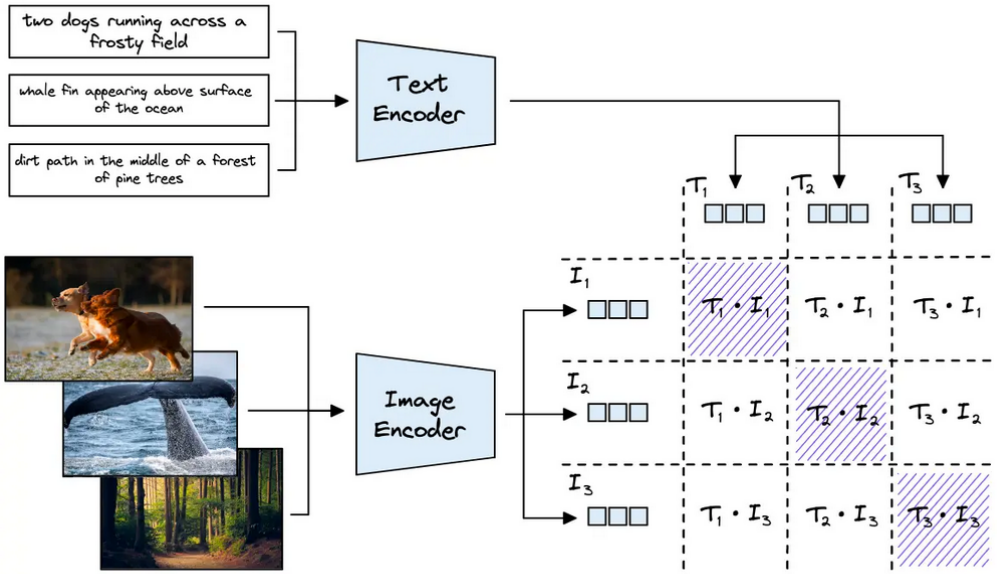
- Pretrained on a dataset of internet image-text pairs
- Two separate encoders: one for images (Vision Transformer) and another for text (Large Language Model).
- Contrastive Learning: The model learns by maximizing the similarity between image and text embeddings for matching pairs



# CLIP: Contrastive Language-Image Pre-Training

## Main Ingredients

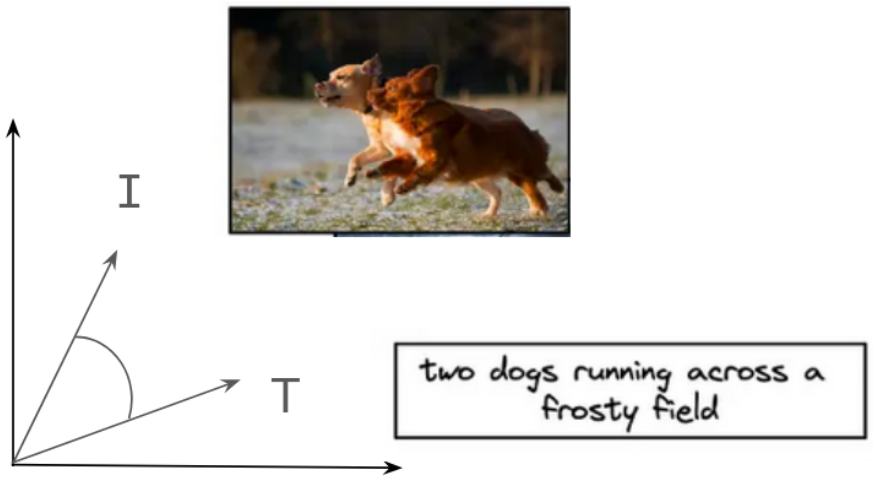
- Pretrained on a dataset of internet image-text pairs
- Two separate encoders: one for images (Vision Transformer) and another for text (Large Language Model).
- Contrastive Learning: The model learns by maximizing the similarity between image and text embeddings for matching pairs



# Image-Text Contrastive Learning Pretraining

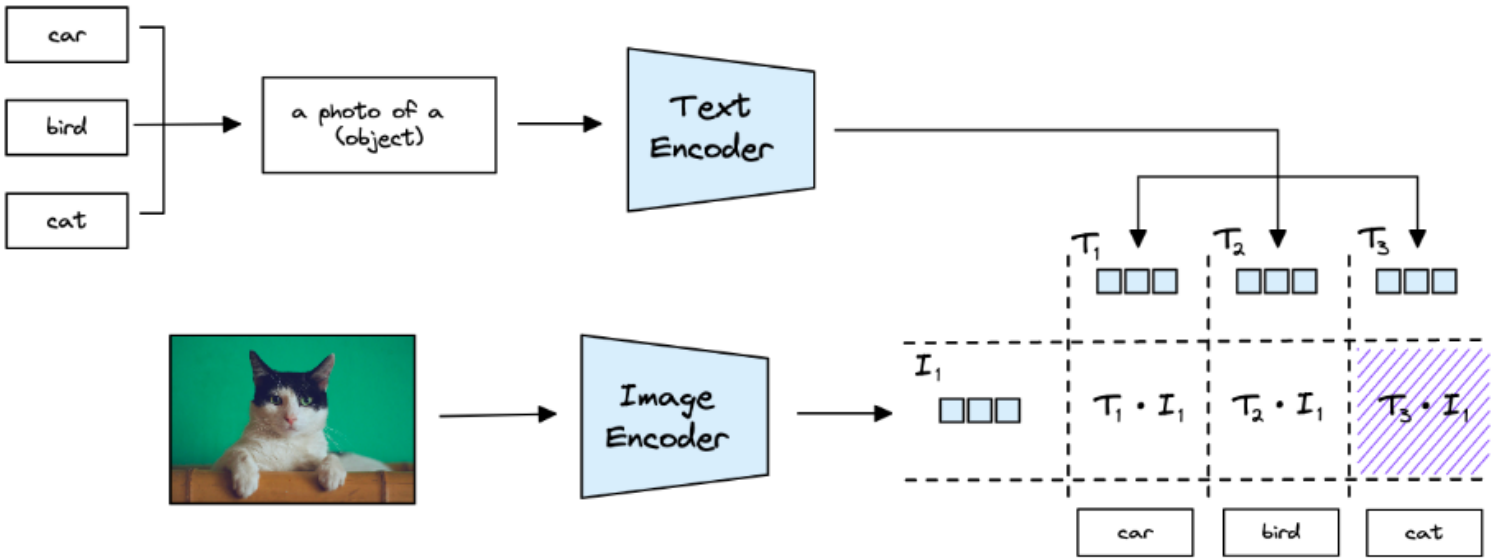
- Embeddings of Image and text Encoders are normalized $||X|| = 1$
- Cosine similarity between normalized features is computed $sim(I, T) = \frac{I \cdot T}{||I||||T||}$
- Maximize similarity of text and images embeddings of same class

$$\mathcal{L}_{CLIP} = -\frac{1}{2N} \sum_{i=1}^{N} \left[ \log \frac{e^{\text{sim}(T_i, I_i)/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(T_i, I_j)/\tau}} + \log \frac{e^{\text{sim}(I_i, T_i)/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(I_i, T_j)/\tau}} \right]$$

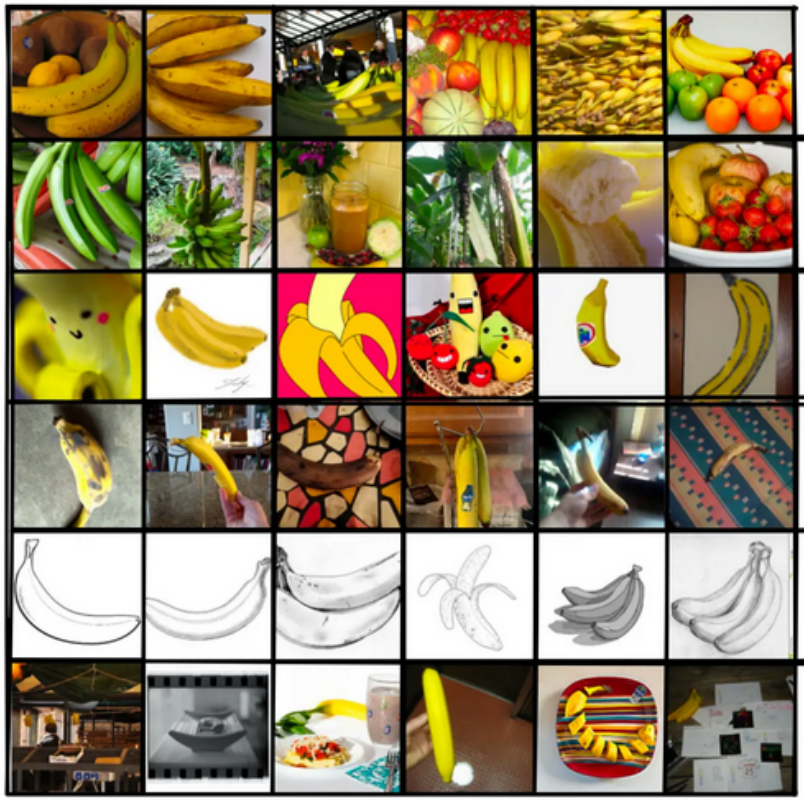


# CLIP Performances

- Zero Shot Classification
- Image Retrieval
- ...



# CLIP Performances

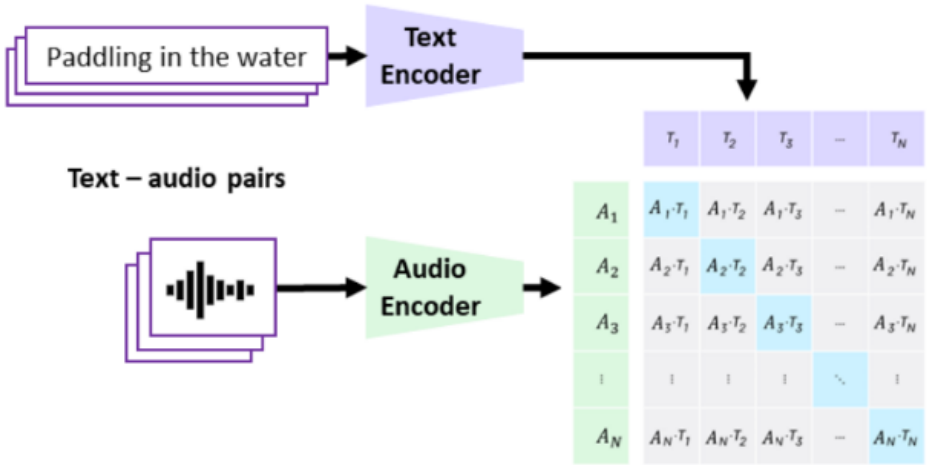


| Dataset Examples | ImageNet ResNet101 | Zero-shot CLIP | Δ Score |
|---|---|---|---|
| ImageNet | 76.2 | 76.2 | 0% |
| ImageNetV2 | 64.3 | 70.1 | +5.8% |
| ImageNet-R | 37.7 | 88.9 | +51.2% |
| ObjectNet | 32.6 | 72.3 | +39.7% |
| ImageNet Sketch | 25.2 | 60.2 | +35.0% |
| ImageNet-A | 2.7 | 77.1 | +74.4% |

# CLAP: Contrastive Language-Audio Pretraining

## Main Ingredients

- Pretrained on 128k audio and text pairs
- Two separate encoders: one for audio (*e.g.*, CNN14) and another for text (*e.g.*, BERT).
- Contrastive Learning: The model learns by maximizing the similarity between image and text embeddings for matching pairs




*CLAP:* `https://arxiv.org/abs/2206.04769`

# CLAP: Contrastive Language-Audio Pretraining

## Main Ingredients

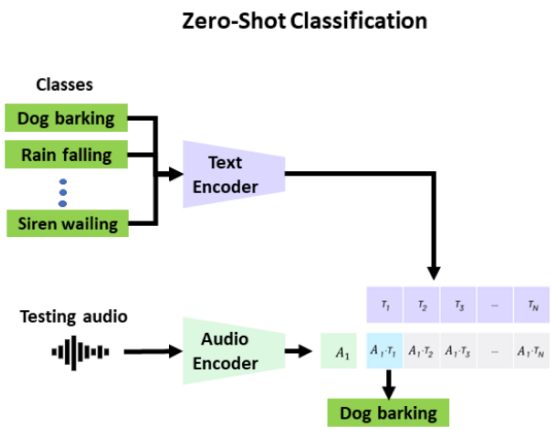
- Pretrained on 128k audio and text pairs
- Two separate encoders: one for audio (*e.g.*, CNN14) and another for text (*e.g.*, BERT).
- Contrastive Learning: The model learns by maximizing the similarity between image and text embeddings for matching pairs

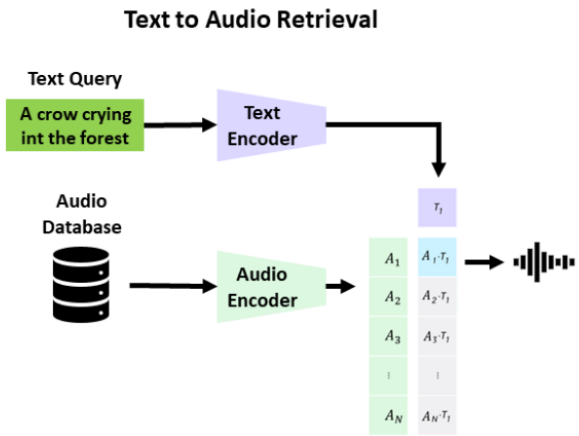
## Different CLAP models

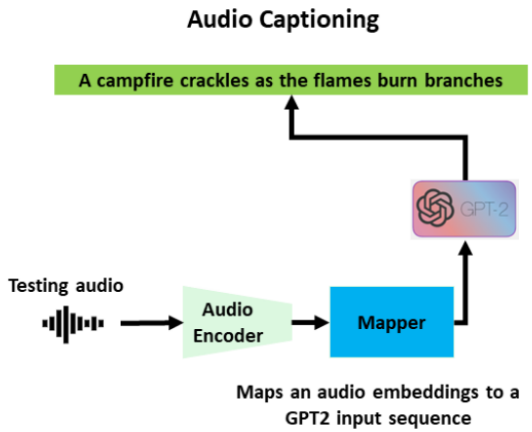
Different versions have been pretrained, all with a similar contrastive approach

- MS-CLAP , Elizalde et al. 2022
- LAION-CLAP, Wu et al. 2023
- MS-CLAP with captioning, Elizalde et al. 2023
- WavCaps, Meil et al. 2023

# CLAP Performances

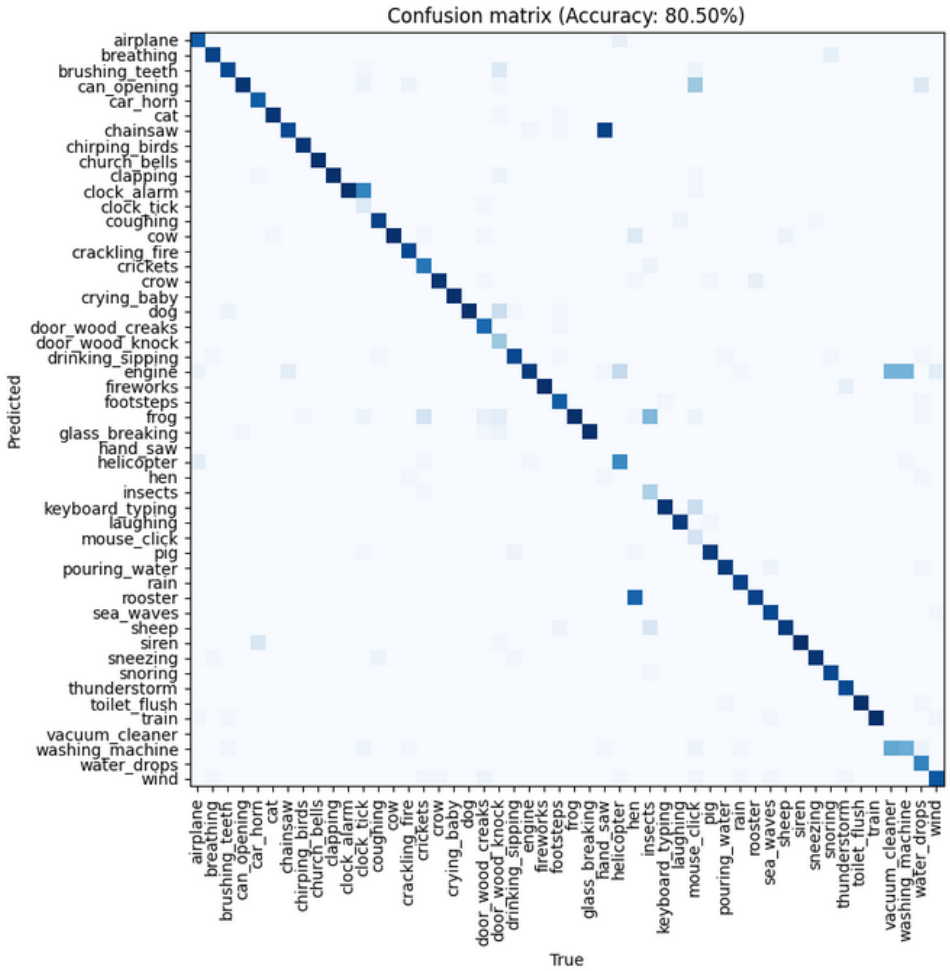


# CLAP Performances

- Example of application to ESC-50 : 50 classes of environmental sounds, 2000 samples, 5 seconds each.




*CLAP:* `https://github.com/jonathanlys01/DL_2023_CLAP`