CLIP Project

“Learning Transferable Visual Models From Natural Language Supervision” by OpenAI

Team:
- Imad Eddine MAROUF
- Anton EMELCHENKOV

Supervisors: - Prof. Marc Lelarge
- Dr. Andrei Bursuc

Agenda

01 Introduction
CLIP Model, OpenAI

02 Experiments
- Zero-shot learning
- Image Retrieval with CLIP as feature extractor

03 Conclusion
Conclude our works

04 Future Work

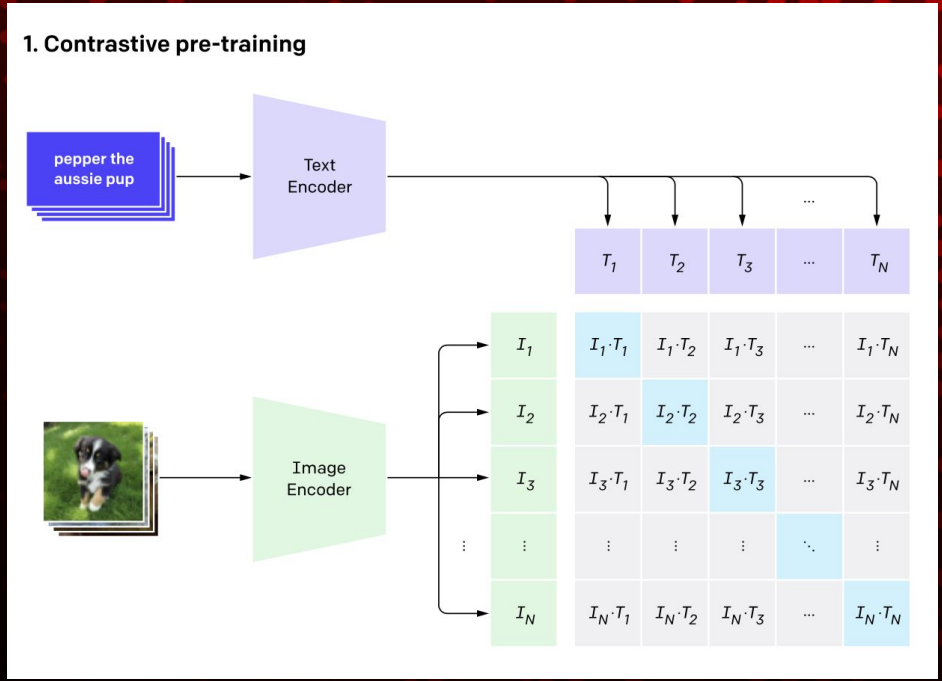
01 Introduction

CLIP Model Architecture,
and One-Shot Learning

CLIP: main concepts

Matches text & images:

- Combines recent advances in NLP and CV.
- Encode image and text and create a similarity matrix for each of text and image classes
- CLIP is trained to predict how likely given image correspond to the text.
- Trained on a very large dataset (400+ million pairs of (text, image) with batches of size n = 32768



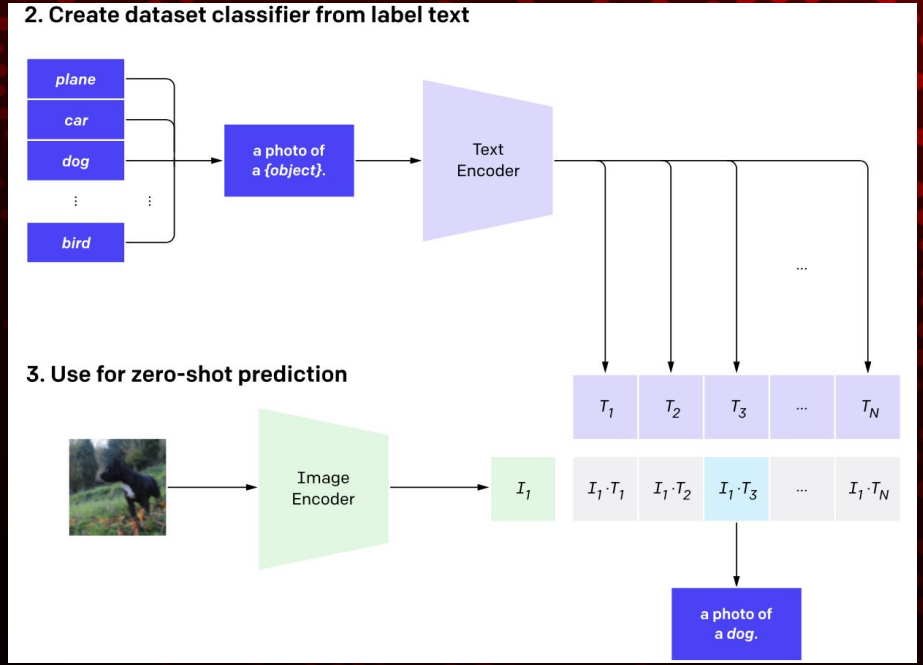
Training the model

Source: OpenAI blog

CLIP: main concepts

Zero-shot prediction:

- CLIP can predict on the classes not observed in training
- Create a set of classes with text descriptions.
- Encode image and text to create a similarity vector for each of the text classes.



Prediction stage

Source: OpenAI blog

CLIP: pros and cons

Pros:

- Zero-shot learning classifier: no need to fine-tune the model
- Can be used on any dataset
- Responds to drawings and text on images
- State-of-the-art model
- Trained on unfiltered, highly varied and noisy data

Cons:

- Sensitive to words and phrasing
- To enhance performance of the model manually annotated descriptions are recommended
- Weak generalization ability on images not covered in its pre-training dataset

02 Experiments

Experiment #1: Zero-Shot Learning: Oxford Buildings

Sample images from Oxford buildings dataset

Dataset description:
- 17 classes: buildings and street view of Oxford
- Relatively hard to match text description with the images.

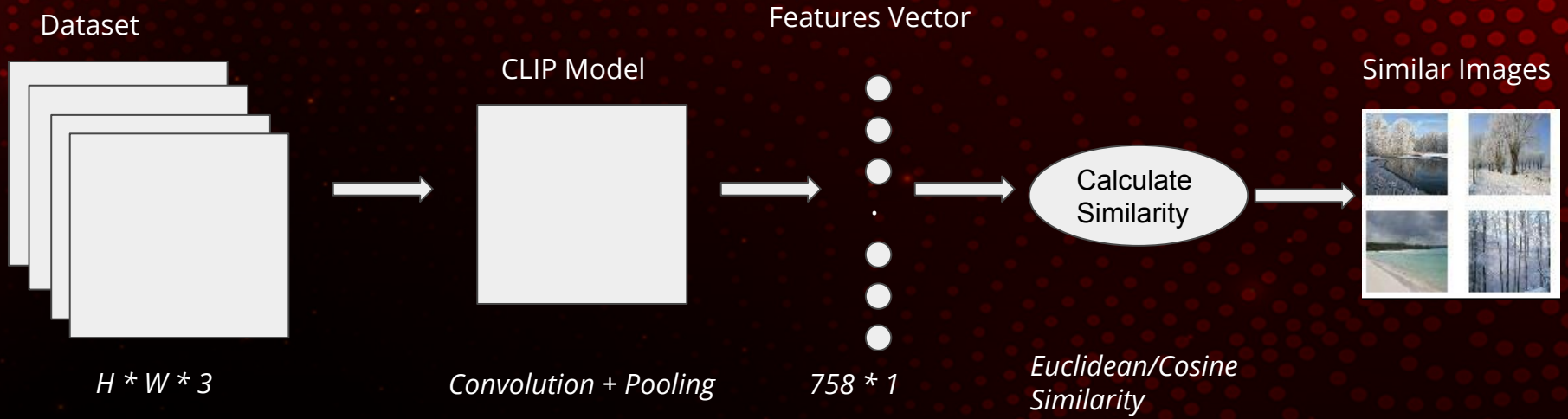
Text prompt:

1 - *"This is a photo of " + class label*
Accuracy: *Top 1*: 19.9%, *Top 5:* 57.1%

2 - *Raw Wikipedia description*
Accuracy: *Top 1*: 27.5%, *Top 3:* 68.3%

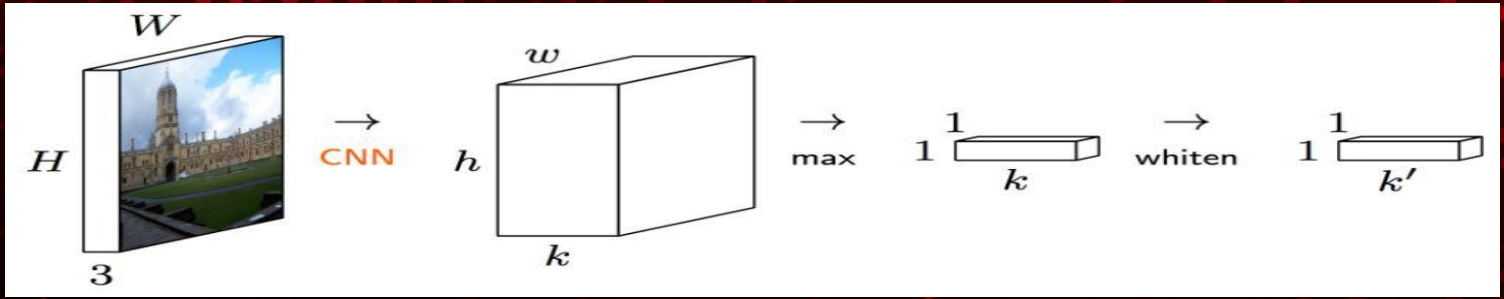
3 - *Manually processed Wikipedia description*
Accuracy: *Top 1*: 31.1%, *Top 3:* 70.5%

Experiment #2: Image Retrieval

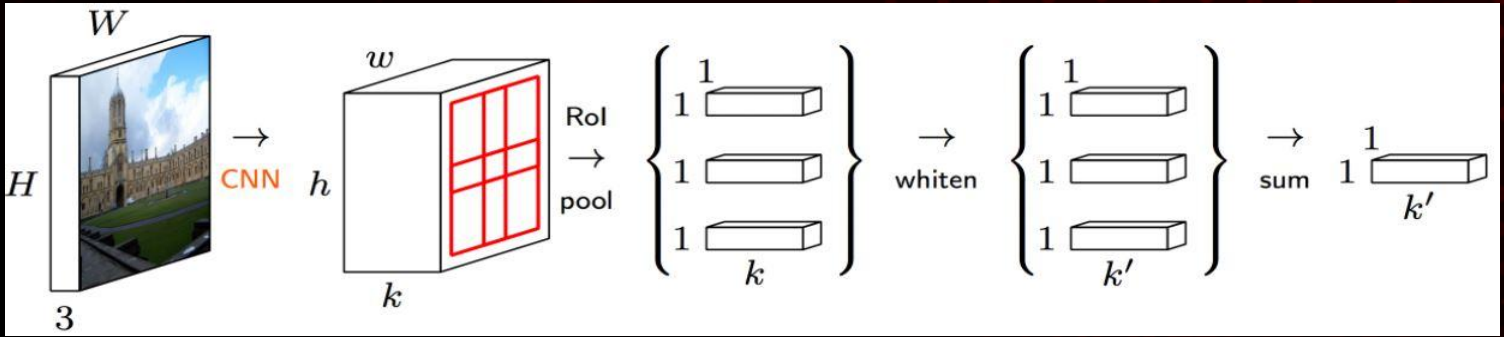


Experiment #2: Image Retrieval

Integral max-Pooling



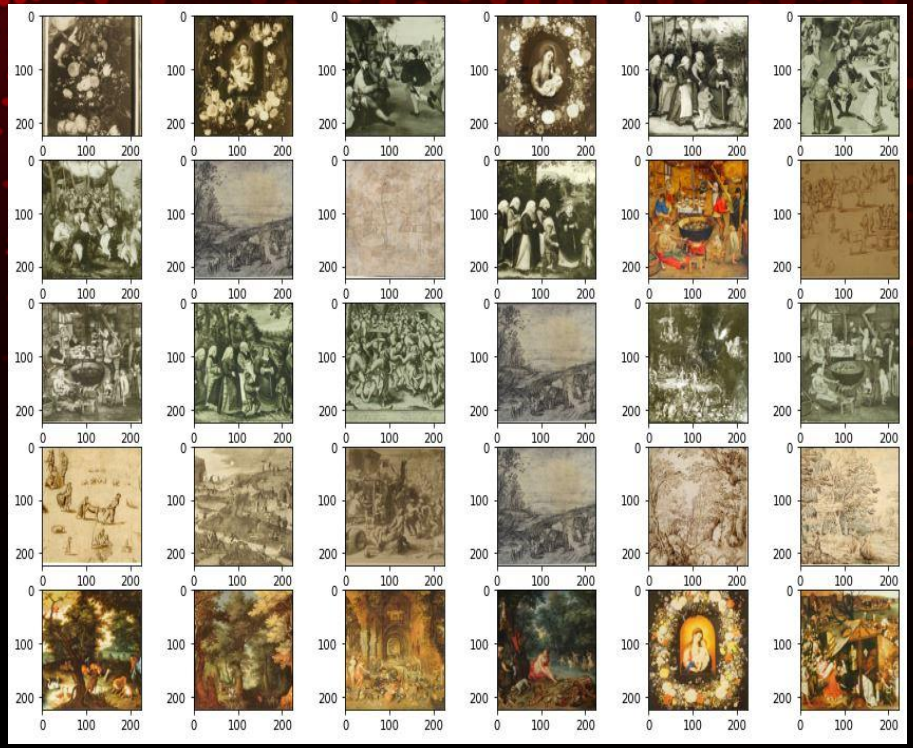
Regional max-Pooling



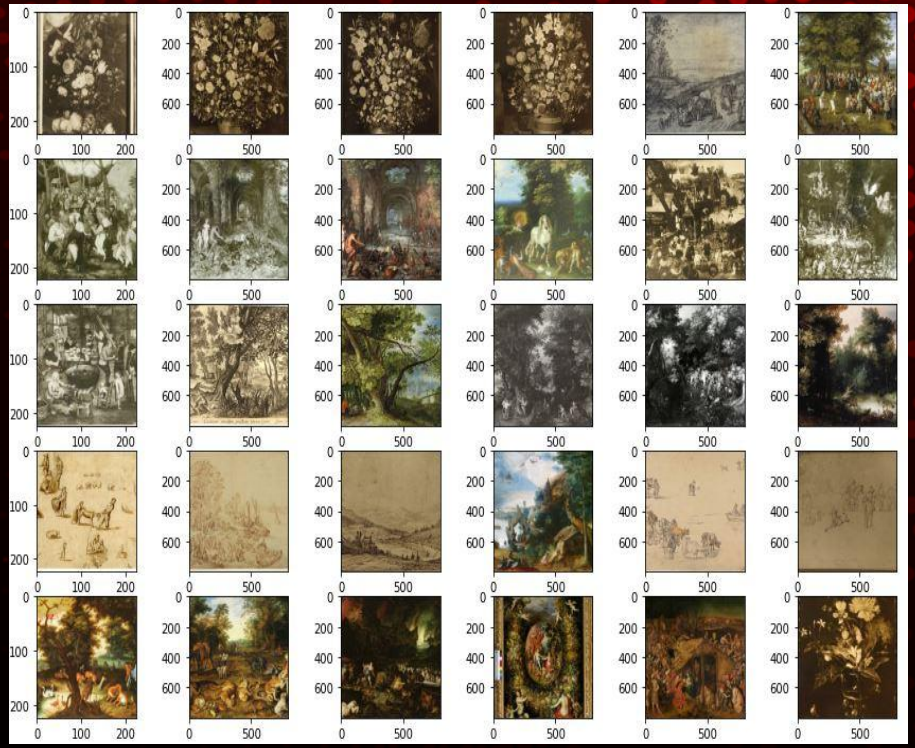
Source: https://arxiv.org/pdf/1511.05879.pdf

Experiment #2: Image Retrieval on Paintings Dataset

Integral max-Pooling

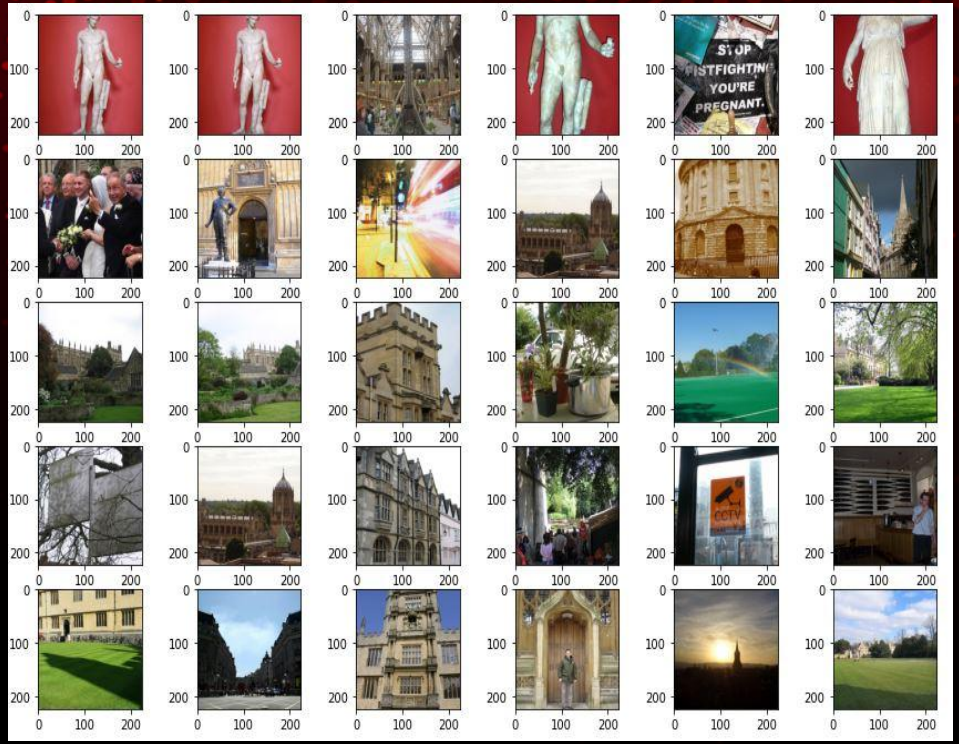


Regional max-Pooling (5 Patches with 40% overlap)

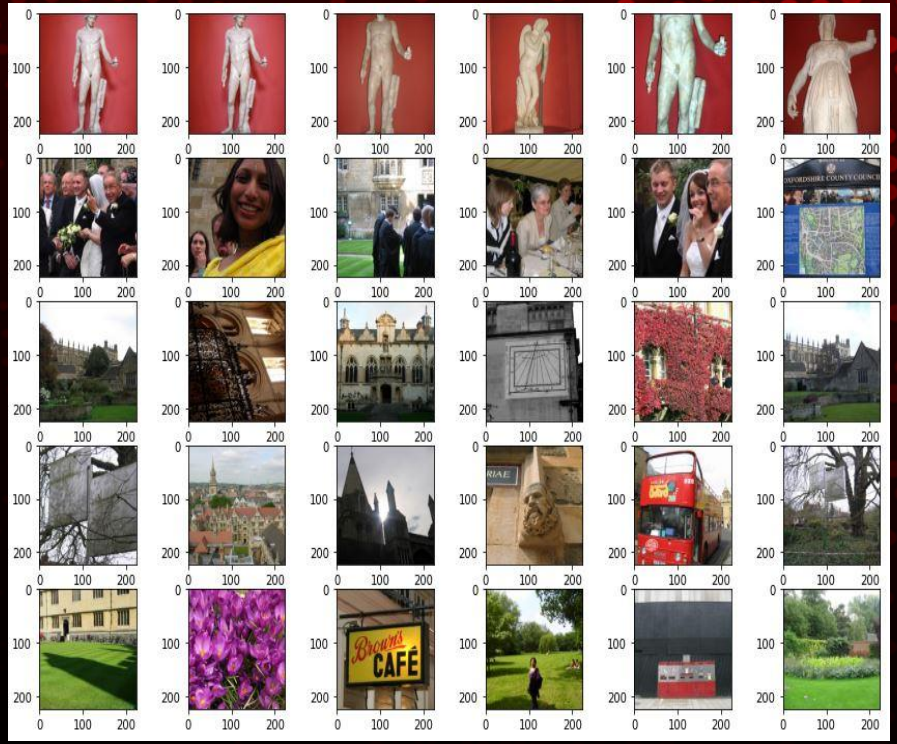


Experiment #2: Image Retrieval on Oxford Dataset

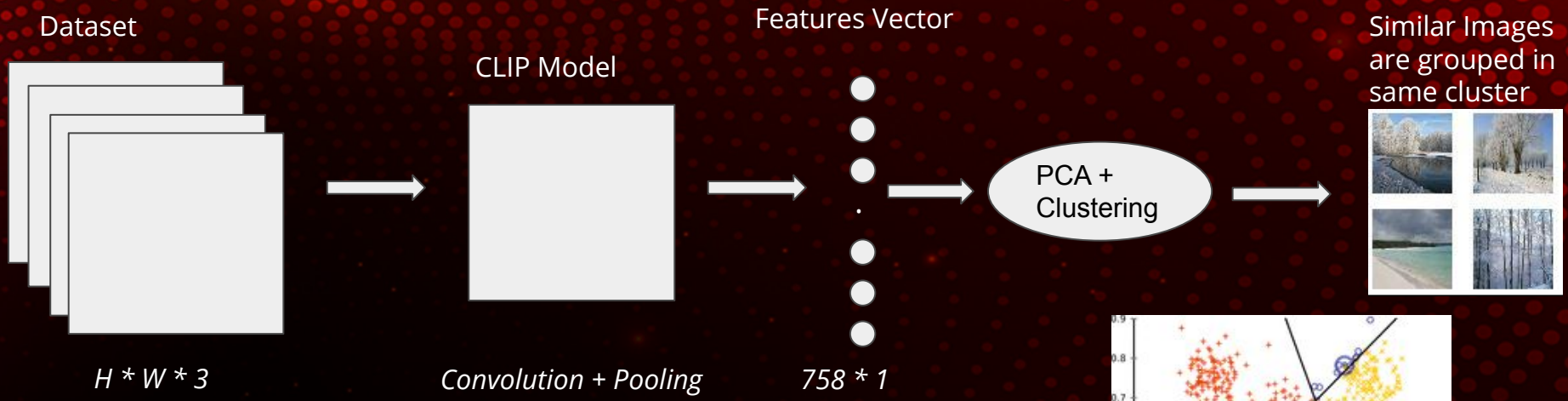
Integral max-Pooling



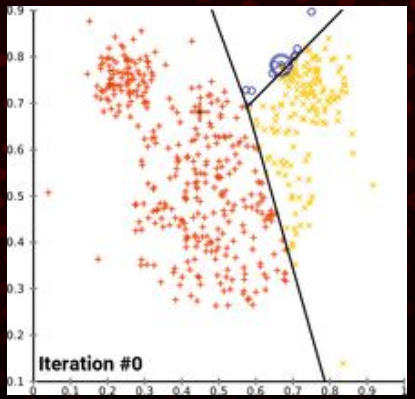
Regional max-Pooling (5 Patches with 40% overlap)



Experiment #2: Improve Image Retrieval

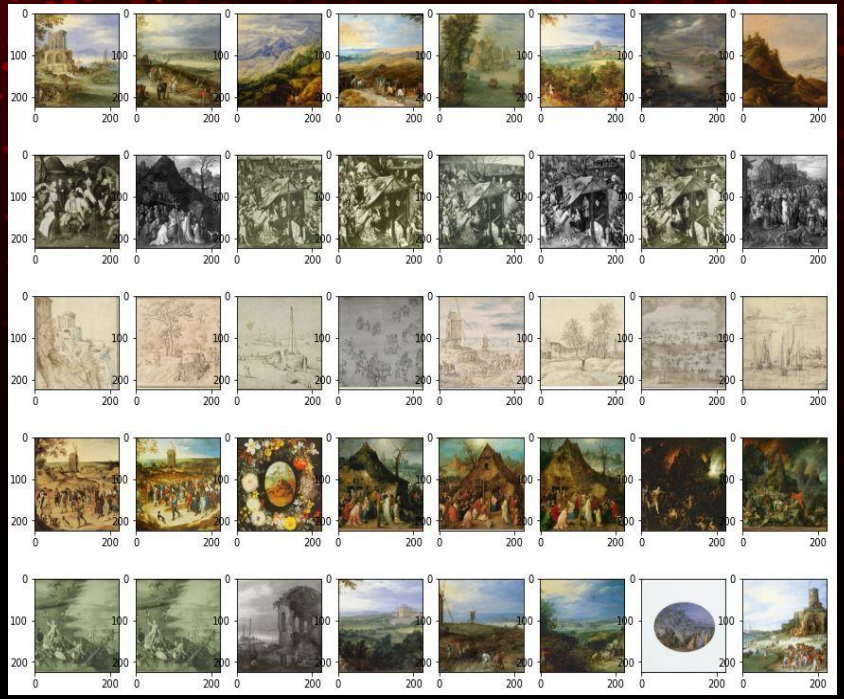


- Principal Component Analysis to [1D * 100]
  - Reduce computational complexity
- Clustering with K-means algorithm (Unsupervised manner)

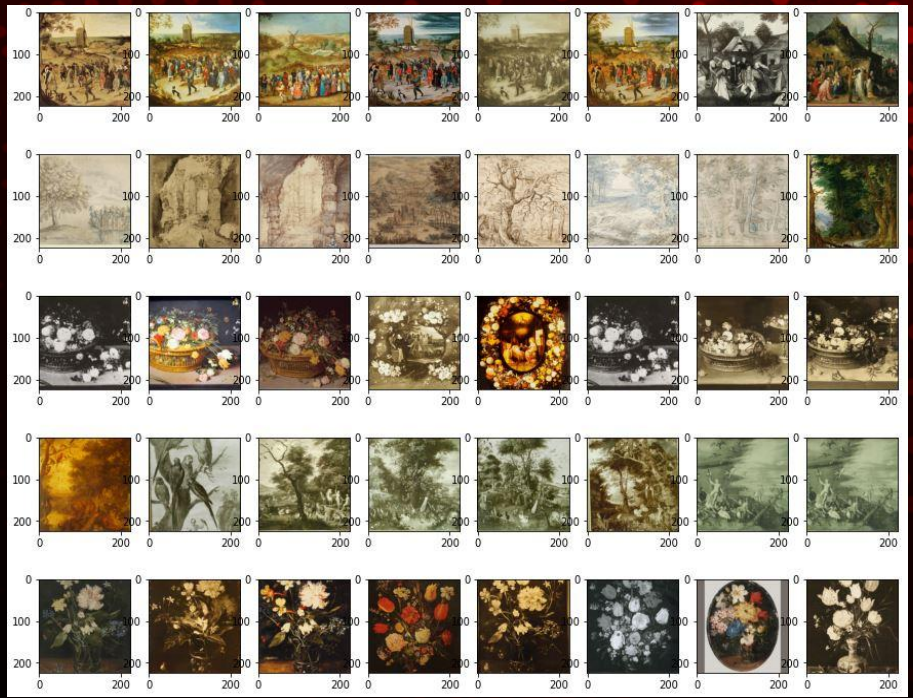


Experiment #2: Image Retrieval on Paintings Dataset

Using only convolutional layers

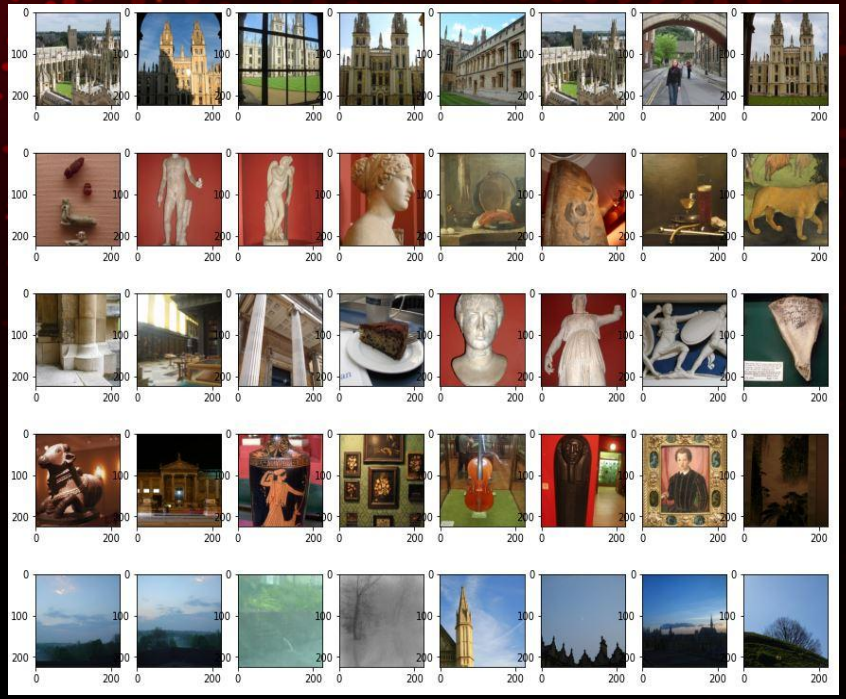


Using all visual network of CLIP

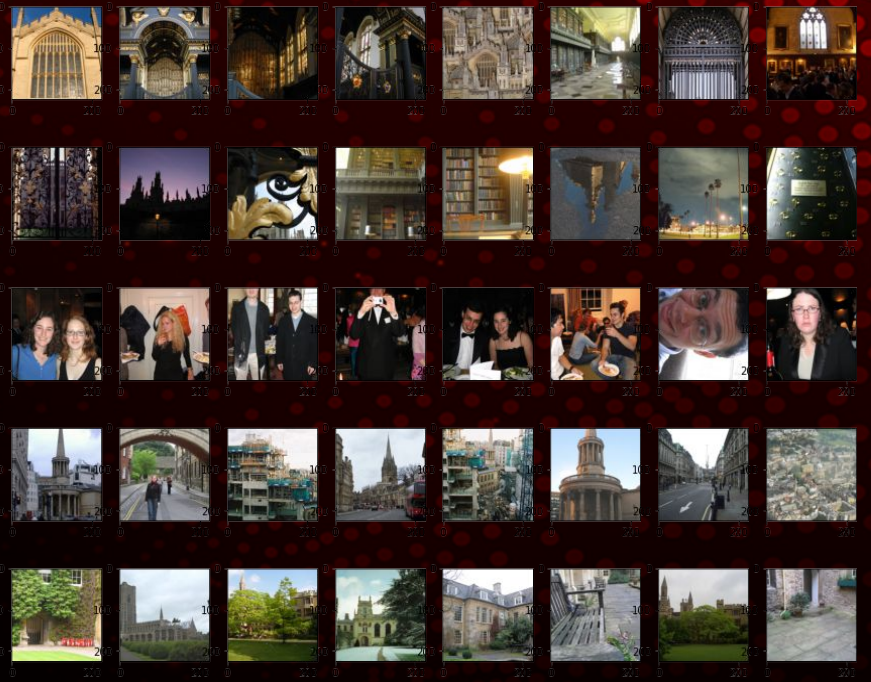


Experiment #2: Image Retrieval on Oxford Dataset

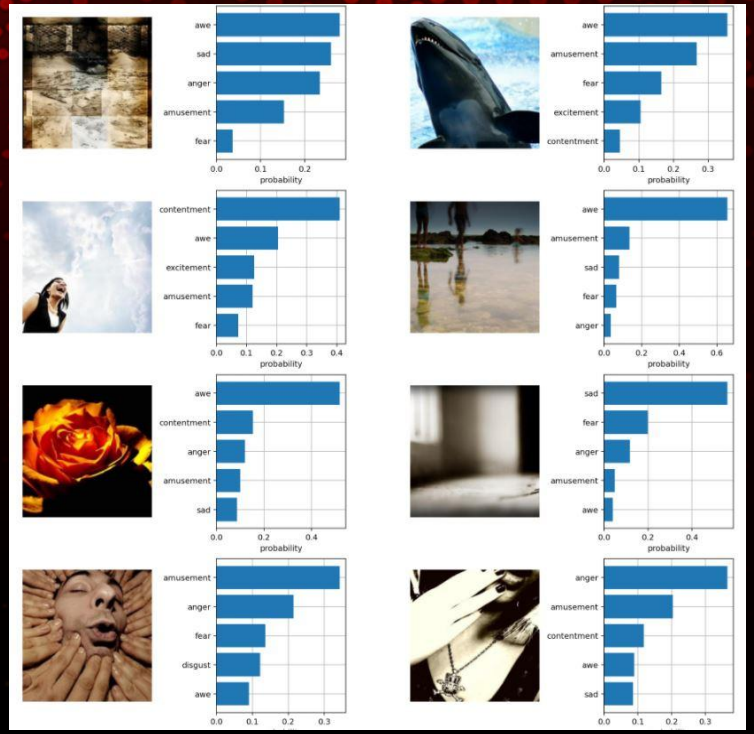
Using only convolutional layers



Using all visual network of CLIP



Experiment #3: *Zero-Shot Learning on Emotions Dataset*



Classes:

- Amusement, Anger, Awe, Contentment, Disgust, Excitement, Fear, Sad

Text prompt:

1 - *"This is a photo of ...."*
Accuracy: *Top 1*: 32.5%, *Top 5:* 78.3%

2 - *Cambridge Dictionary definition*
Accuracy: *Top 1*: 27.5%, *Top 3:* 57.7%

3 - *"Name" + "is " + Dictionary definition*
Accuracy: *Top 1*: 28.0%, *Top 3:* 57.8%

Linear-Probe evaluation:

- Accuracy: Training: 95.5%, Test: 45.631%

03 Conclusion

Conclusion

- Transfer learning is very powerful technique as a baseline.
  - No need to reinvent the wheel (Use off-the-shelf networks)

- Oxford Dataset/Brueghel Paintings are very challenging.
  - Noisy and low resolution.
  - Hard to distinguish between images from different clusters.

- Distance metric is important.
  - Cosine similarity is efficient, but was not good for our experiments.

04 Future Work

Future Work

- Fix PCA Whitening for feature extraction.

- Metric to measure models performance on image ranking.

- Experiment: Typographic attacks on Oxford/Emotions datasets
  - *Typographic attacks: adding adversarial text to image can cause them to be systematically misclassified*

    Source: Multimodal neurons in Artificial Neural networks (https://distill.pub/2021/multimodal-neurons/)

?????????

Thank you for your attention

??? Questions ?

?????????????

?????????